illumina®

Illumina DRAGEN Bio-IT Platform v3.9

User Guide

ILLUMINA PROPRIETARY
文書番号：200005495 v00 JPN
2021年7月

本製品の使用目的は研究に限定されます。診断での使用はできません。

本製品の使用目的は研究に限定されます。診断での使用はできません。

目次

はじめに

始める前に、Illumina® DRAGEN™ Bio-IT Platformサーバーの電源が入っていること、サーバーにログインしていることを確認してください。

この「はじめに」セクションは、ユーザーができるだけ速やかにデータの処理を開始できるよう支援することを目的に、以下についての手順を説明します：

DRAGENには、使用しているDRAGENシステムが適切に設置、構成されていることを確認するためのテストが用意されています。このテストを実行する前に、DRAGENサーバーに十分な電源が提供され、適切に冷却されていること、およびDRAGENサーバーとの間でデータを適切なパフォーマンスでやりとりするに十分な速度を持つネットワークに接続されていることを確認してください。

DRAGEN Bio-It Platformの概要

Illumina DRAGEN™ Bio-IT Platformは、再構成しやすいDRAGEN Bio-IT Processorをベースにしています。このプラットフォームは、Field Programmable Gate Array(FPGA)カードに一体化されていて、バイオインフォマティクスワークフローへシームレスに統合可能な事前構成済みのサーバーで使用することができます。このプラットフォームには、以下を対象とした多種多様なNGS二次解析パイプラインに対して高度に最適化されたアルゴリズムをロードできます：

- 全ゲノム
- エクソーム
- RNA-Seq
- メチローム
- がん

対話式操作はすべて、DRAGENソフトウェア経由で行われます。このソフトウェアは、ホストサーバー上で実行され、DRAGENボードとのコミュニケーションすべてをつかさどります。

このユーザーガイドでは、本システムの技術面を簡潔にまとめて説明し、DRAGENの全コマンドラインオプションについて詳しく解説します。

DRAGENを使うのが初めての場合は、まず、1 ページの 「はじめに」 セクションを参照することを推奨しています。このセクションでは、サーバーテストの実行、リファレンスゲノムの生成、サンプルコマンドの実行など、DRAGENを簡単に説明しています。

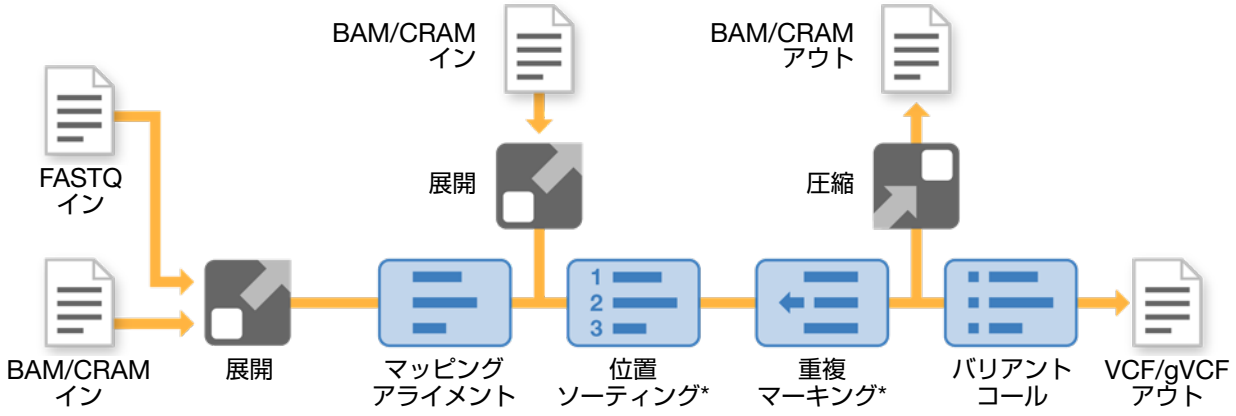
インプット要件

DRAGENでサポートされているインプット要件は以下のとおりです。

仕様	要件
サポート対象の入力ファイル	cBCL、FASTQ、BAM、CRAM、GVCF
上限	ヒト全ゲノムシーケンスデータ300xカバレッジ。ヒトT/N 200x/100xカバレッジ。

DRAGEN DNA Pipeline

図 1　DRAGEN DNA Pipeline



*オプション

DRAGEN DNA PipelineはNGSデータの二次解析を加速します。例えば、30xカバレッジでヒトゲノム全体を処理する際にかかる時間が、約10時間（現在の業界標準であるBWA-MEM+GATK-HCソフトウェアを使用）から約20分に短縮されます。所要時間は、カバレッジ深度に対して直線的に増減します。

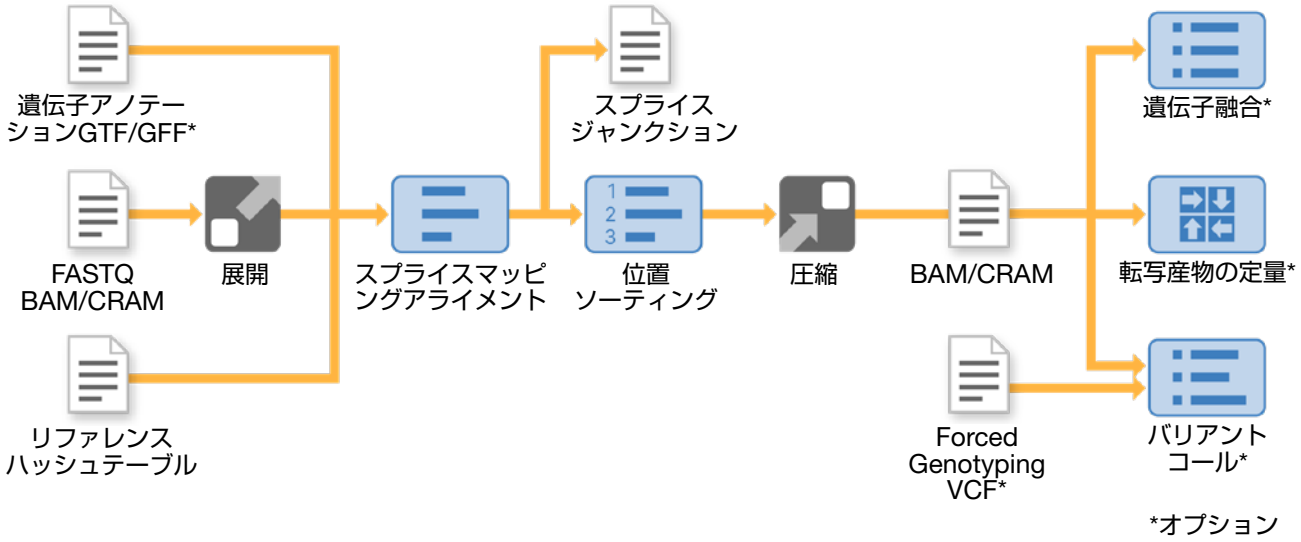
これらのパイプラインは、DRAGEN Bio-It Platformの大きな力を活用し、マッピング、アライメント、ソーティング、重複マーキングおよびハプロタイプバリアントコールを行うための高度に最適化されたアルゴリズムを搭載しています。また、プラットフォームが備えているあらゆるツールとともに、ハードウェアアクセラレーションによる圧縮機能、最適化されたBCL変換など、プラットフォームの機能も使用します。

他の二次解析メソッドとは異なり、DRAGENのDNAアプリケーションでは、速度が向上したからといって、精度が低下することはありません。BWA- MEM+GATK-HCと対照比較した場合、SNP、INDELともに精度が向上しています。

このパイプラインは、ハプロタイプバリアントコールに加え、コピー数バリアントや構造多型のコール、リピート伸長の検出も対応しています。

DRAGEN RNA Pipeline

DRAGENには、RNA-Seq（スプライシング対応）アライナーと、遺伝子発現の定量および遺伝子融合検出のためのRNA固有の解析コンポーネントがあります。



DRAGEN RNA Pipelineは、DNA Pipelineと多数のコンポーネントを共有しています。RNA-Seqリードからの短いシードシーケンスのマッピングは、DNAリードのマッピングと同じように行われます。さらに、マッピングされたシードに近接するスプライスジャンクション（RNA転写産物中の隣接していないエクソン同士の結合部分）を検出し、完全なリードアライメントに取り込みます。

DRAGEN Methylation Pipeline

DRAGEN Methylation Pipelineは、メチル化解析に必要なタグのついたBAMを生成するためのバイサルファイトシーケンシングデータの処理の自動化に対応し、メチル化されたシトシンの詳しい位置をレポートします。

実行要件

DRAGEN Bio-It Platform Analysisアプリで、解析を完全に実行するには、ランフォルダー内に以下のファイルが必要です：

- BCLファイル（*.bcl、*.cbcl）
- フィルターファイル（*.filter）
- 位置ファイル（*.locs、*.clocs、*s.locs）
- アグリゲートファイル（*.bci）
- 実行情報ファイル（*.xml）
- Config.xml：config.xmlファイルは、一部のシステムで作成されたデータでのみ必要です。詳細については、イルミナサポートサイトのDRAGENに関するページを参照してください。

シーケンスデータ

DRAGENを実行するには、以下のファイルが必要です。データの作成に使われているシーケンスシステムによっては、異なるインプットを要求されることがあります。

インプット	説明
ベースコールファイル（*.bcl.bgzf）	ベースコール（BCL）ファイルは、gzip（*.gz）形式で圧縮されているか、またはGNU zip（*.bgzf）形式でブロック化されています。
ベースコールインデックスファイル（*.bci）	ベースコールインデックス（BCI）ファイルには、1レーンにつき1つの記録がバイナリー形式で含まれています。BCIファイルは、DRAGENによるインプットとしては許容されますが、解析には使用されません。
連結ベースコールファイル（*.cbcl）	連結ベースコール（CBCL）ファイルには、集計されたBCLデータが含まれます。同じレーンとサーフェスのタイルが、レーンとサーフェスごとに1つのCBCLファイルにまとめられます。
フィルターファイル（*.filter）	フィルターファイルは、所定のクラスターがフィルターを通過するかどうかを規定したバイナリーファイルです。
位置ファイル（*.locs、s.locs）	位置ファイル（LOCS）は、フローセル上でのクラスターの位置を含むバイナリーファイルです。CLOCSファイルは、LOCSファイルの圧縮版です。
実行情報ファイル（RunInfo.xml）	実行情報ファイルは、アウトプットフォルダーのルートレベルにあります。このファイルには、ラン名、サイクル数、リードがインデックスリードかどうか、レーン、スワス、タイルの数が含まれます。アウトプットフォルダーにこのファイルが存在しない場合、ソフトウェアはエラーになります。
構成ファイル（*.xml）	構成ファイルはBaseCallsフォルダーにあり、シーケンスランのメタデータが入っています。このファイルはXML形式です。

DRAGENの実行

BCL変換でDRAGENを実行する場合は、以下の情報を使用します。

コマンドライン経由でBCL変換を実行するときに切断や端末の終了を防ぐには、コマンドを実行する前に、`nohup`と入力します。

ulimit設定

DRAGENでは、一度に開くことができるファイルの数と最大ユーザープロセス数の両方の上限（ulimit）を高く設定する必要があります。最大ユーザープロセス数が小さすぎたことが原因で実行が失敗した場合、一時的にリソースが利用できなくなったというエラーメッセージが表示されます。初期設定では、DRAGENは、開くことができるファイル数のulimitソフト制限（ulimit-n）を65535、最大ユーザープロセス数を32768に設定しようと試みます。これらの値が、システムのハード制限を上回る場合、ソフト制限にはハード制限の値が設定されます。

提供されたサンプルの数が10,000を上回る場合、`ulimit -n`は、`720000`に設定されます。

不明ファイルへの対応

`--strict-mode`がfalseに設定されている場合、DRAGENは不明または破損したファイルが見つかったときに、特定の動作を行います。実行される可能性のある動作は、以下のように、ファイルの種類と状況によって異なります。

ファイルの種類	状況	動作
*.bcl	不明または破損	該当のレーンとタイルにあるサイクルの全ベースコールを、Nとクオリティスコア#で置き換えます。
*.cbcl	不明または破損	該当のレーンとサーフェスにあるサイクルの全ベースコールを、Nとクオリティスコア#で置き換えます。
*.cbcl	破損	該当のレーンとタイルにあるサイクルの全ベースコールを、Nとクオリティスコア#で置き換えます。
*.locs	不明または破損	該当のレーンとタイルにある全リードに対して自動的に生成された固有ヘッダーを使って、FASTQファイルを作成します。
*.filter	不明または破損	該当のレーンとタイルにあるリードに対して作成されるFASTQエントリーはありません。
*.bci lane	不明または破損	該当のレーンとタイルにあるリードに対して作成されるFASTQエントリーはありません。

ソフトウェアのインストール

DRAGENソフトウェアとハードウェアの最新バージョンをすでに使用されている場合は、7 ページの「システムチェックの実行」 に進みます。

現在のソフトウェアとハードウェアのバージョンは、以下のコマンドを使って調べることができます：

```
dragen_info -b
```

ソフトウェアのバージョンだけを調べるには、以下のコマンドを使います：

```
dragen --version
```

新しいバージョンのソフトウェアまたはハードウェアをインストールするには、まず、イルミナのウェブサイトにあるDRAGEN Bio-IT Platformのサポートページからサーバーへパッケージをダウンロードします。インストールメソッドとしては、以下のように、自己解凍式の.runファイルを使用することを推奨します：

```
sudo sh dragen-3.3.7.el7.x86_64.run
```

インストール中に、新しいバージョンのハードウェアへの切り替えを勧めるメッセージが表示された場合は、「y」と入力します。ハードウェアのアップグレードプロセスが中断されないようにすることが重要です。完了したら、サーバーを停止し、一度電源を切って、再度立ち上げなおします。リブートコマンドでは、ハードウェアのバージョンは更新されません。サーバーの電源をオフにして、再びオンにするには、次のhaltコマンドを使用する必要があります：

```
sudo ipmitool chassis power cycle
```

DRAGENは、ライセンスの更新状況をチェックするため、定期的にlus.edicogenome.comにあるライセンスサーバーと通信します。ファイアウォールの背後にあるサーバーについては、ネットワーク管理者が/etc/environmentにプロキシを設定できます。例えば：

```
http_proxy="http://proxy.customer.com:80/"
https_proxy="https://proxy.customer.com:80/"
ftp_proxy="http://proxy.customer.com:80/"
rsync_proxy="http://proxy.customer.com:80/"
no_proxy="localhost,127.0.0.1,localaddress,.localdomain.com,.customer.com"
```

システム更新

DRAGENサーバーにあるソフトウェアはすべて更新できます。DRAGEN Bio-It Platformサポートサイトからソフトウェアの新バージョンをダウンロードできます。

カーネルパッケージは、CentOS Updatesからのみ入手可能です。実験用カーネルを使用したり、ソースからカーネルをコンパイルしたりすることは推奨しません。

ライセンスの使用状況

現在のライセンスの使用状況と有効期限を調べるには、次のコマンドを使用します：

```
dragen_lic -f genome
```

このコマンドにより、次のようなライセンス情報が出力されます：

```
LICENSE_MSG| ---- Board #0 (1234565) ----
LICENSE_MSG| License Genome: used 1000.0/100000 Gbases since 2019-Jan-01
(1000000000000 bases, 1.0%)
LICENSE_MSG| Issued=2019-Jan-01, start=2019-Jan-01, expiry=2020-Jan-01,
period=12 months


LICENSE_MSG| -- License dongle
LICENSE_MSG| STATUS : OK
LICENSE_MSG| DONGLE SN: 0012345678900
LICENSE_MSG| RELEASE : 2016.07p5-19358
LICENSE_MSG| CHIPID : 001234567890EAD
LICENSE_MSG| DNA: active, accelerators=DNA
LICENSE_MSG| issue=2019-Jan-01, start=2019-Jan-01, expiry=2020-Jan-01
LICENSE_MSG| RNA: active, accelerators=RNA
LICENSE_MSG| issue=2019-Jan-01, start=2019-Jan-01, expiry=2020-Jan-01
```

```
LICENSE_MSG| GZIP: active, accelerators=GZIP
LICENSE_MSG| issue=2019-Jan-01, start=2019-Jan-01, expiry=2020-Jan-01
LICENSE_MSG| GUNZ: active, accelerators=GUNZ
LICENSE_MSG| issue=2019-Jan-01, start=2019-Jan-01, expiry=2020-Jan-01
LICENSE_MSG| HMM: active, accelerators=HMM
LICENSE_MSG| issue=2019-Jan-01, start=2019-Jan-01, expiry=2020-Jan-01
LICENSE_MSG| SMW: active, accelerators=SMW
LICENSE_MSG| issue=2019-Jan-01, start=2019-Jan-01, expiry=2020-Jan-01
LICENSE_MSG| RANS: active, accelerators=RANS
LICENSE_MSG| issue=2019-Jan-01, start=2019-Jan-01, expiry=2020-Jan-01
LICENSE_MSG| GRAPH: active, accelerators=GRAPH
LICENSE_MSG| issue=2019-Jan-01, start=2019-Jan-01, expiry=2020-Jan-01
```

前述のライセンス出力例は、ゲノムライセンスのものです。先頭行は、既に1,000ギガベースが使用されていることを示しています。インストール済みのライセンスは100,000ギガベース用で、そのうち1%が使用済みです。2行目はライセンスデータを示しています。ここで重要なのは有効期限です。ライセンスは、有効期限の期日、またはライセンスされたギガベースが100%使用されたときに失効します。

ライセンスデータの下に続くのは、ライセンス情報です。この情報は、サーバーに接続されたドングル（USBキー）に保存されています。ライセンス情報には、有効化されている全アクセラレーターの状況が表示されます。これらはそれぞれ異なるパイプライン特有のものです。また、各パイプラインのライセンスや、ゲノムライセンスの例と同様、アクセラレーターにも有効期限があります。

新しいライセンスの取得については、カスタマーサービス担当者（customerservice@illumina.com）にお問い合わせください。ライセンスの使用で問題が発生した場合は、イルミナのテクニカルサポートにお問い合わせください。

システムチェックの実行

DRAGENサーバーの電源を入れたら、サーバーが正しく動作することを確認するために`/opt/edico/self_test/self_test.sh`を実行します。このスクリプトは以下を行います。

- hg19リファレンスゲノムからM染色体のインデックスを自動的に作成
- リファレンスゲノムとインデックスをロード
- リードのセットをマッピングし、アライメント
- アライメントしたリードをBAMファイルに保存
- アライメントが予想された結果と完全一致していることを確認

各サーバーは、このスクリプトで使用されるテストインプットのFASTQデータを搭載しています。このデータは、`/opt/edico/self_test`にあります。システムチェックの実行には、約25～30分かかります。

次の例は、スクリプトの実行方法と、テストが成功した場合の出力を示しています。

```
[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
```

出力されたBAMファイルが予想された結果と一致しない場合、前述のテキストの最終行に、次のように表示されます：

```
SELF TEST RESULT : FAIL
```

DRAGENサーバーの電源を入れた直後にこのテストスクリプトを実行して、その結果がFAILになった場合は、イルミナのテクニカルサポートにご連絡ください。

お客様独自のテストの実行

DRAGENシステムが期待どおりに動作していることを確認したら、以下のようにして、装置からお客様独自のデータの一部を実行します。

- リファレンスゲノムのハッシュテーブルをロード
- 入力ファイルと出力ファイルの場所を決定
- インプットデータを処理

リファレンスを生成

リファレンスを持っていない場合には、dragen *–build-hash-table*コマンドにFASTAリファレンスファイルの場所を渡して実行し、生成することができます。ハッシュテーブルを構築するときに、パラメーターを指定できます（『*DRAGEN Bio-IT Platform User Guide*』（1000000070494）を参照）。

テストの目的では、サンプルシェルスクリプト、またはこのガイドの例で使用されているコマンドを実行できます。これらの例は、FASTAファイルが`/staging/human/reference/hg19/hg19.fa`に保管されていることを前提としています。必要に応じて、スクリプトやコマンドライン内のパスを正しいディレクトリに変更してください。`/staging/human/reference`とそのサブディレクトリに対するアクセス権を変更する必要があります。

次のようにして、サンプルシェルスクリプトを実行します：

```
/opt/edico/examples/build_hash_table.sh
```

または、次のようなDRAGENのコマンドを実行します：

```
mkdir -p /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149
cd /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
dragen --build-hash-table true --ht-reference
/staging/human/reference/hg19/hg19.fa \
--output-dir /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
--ht-alt-liftover /opt/edico/liftover/hg19_alt_liftover.sam
```

*--ht-alt-liftover*オプションを指定せずにハッシュテーブルを生成した場合、以下に類似したエラーが表示されます（表示内容は使用した.faリファレンスファイルによって異なります）：

```
ERROR: Detected hg19 alternate contigs in reference at:
/staging/hg19fa/hg19.fa
DRAGEN map quality is significantly improved by building a reference with a
liftover file to enable ALT aware mapping. Use the --ht-alt-lifeover option
to specify a liftover file.
You may ignore this error and continue using your existing reference by
adding --ht-alt-aware-validate=false to your command line. However, DRAGEN
map quality will be significantly affected.
Generate the hash table with either the --ht-alt-liftover or the -ht alt-
aware-validate=false option to avoid the error listed above.
```

dragen *--build-hash-table*コマンドはマルチスレッド化されていて、初期設定は8スレッドです。このコマンドの実行には約15分かかります。ランタイムを短縮するには、*--ht-num-threads*オプションを使用します。指定できる値は最大32です（サーバーがサポートしているスレッド数により異なります）。

ハッシュテーブルディレクトリ名には、ハッシュテーブルの構築中に使用された主要なオプションの初期設定値がリストされます。お客様が独自のハッシュテーブルを生成し、それに応じてディレクトリ名を変更する場合、このベストプラクティスに従うことを推奨します。

CNV機能を有効化した場合、ハッシュテーブルの生成には約2時間かかります。

HG19リファレンスの生成

FASTAリファレンスを持っていない場合、次のようにして、UCSCからhg19 FASTAファイルを取得し、1つのhg19.faファイルに結合することができます。

```
mkdir /staging/hg19fa
cd /staging/hg19fa
wget
hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz
tar -zxvf chromFa.tar.gz
cat chr*.fa > hg19.fa
```

次のコマンドを使って、DRAGENハッシュテーブルリファレンスを生成します。

```
mkdir /staging/hg19/
dragen --ht-reference /staging/hg19fa/hg19.fa \
    --output-directory /staging/hg19/ --build-hash-table true \
    --ht-alt-liftover /opt/edico/liftover/hg19_alt_liftover.sam
```

リファレンスゲノムのロード

DRAGENボードのメモリーにロードしたバイナリーリファレンスは、任意の数のインプットデータセットの処理に使用できます。システムを再起動しない限り、または異なるリファレンスハッシュテーブルを使用する必要がない限り、リファレンスを再ロードする必要はありません。

リファレンスは、データを初めて処理するときに、自動的にロードされます。リファレンスゲノムを手作業でボードにロードする場合は、次のシェルスクリプト、またはコマンドを使用します。この例のリファレンスディレクトリは/staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149です。

```
/opt/edico/examples/load_reference.sh
```

または

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

このコマンドは、DRAGENボードのメモリーにバイナリーのリファレンスゲノムをロードします。そこで、このリファレンスは、任意の数のインプットデータセットの処理に使用されます。システムを再起動しない限り、また、異なるリファレンスゲノムに切り替える必要がない限り、リファレンスゲノムを再ロードする必要はありません。リファレンスゲノムは1分以内にロードされます。

DRAGENは、指定されたリファレンスゲノムが既にボードに存在しているかどうかをチェックします。存在している場合、リファレンスゲノムのアップロードは自動的にスキップされます。同じリファレンスゲノムを強制的に再ロードするには、*force-load-reference (-l)*コマンドラインオプションを使用します。

リファレンスゲノムをロードするコマンドは、ソフトウェアとハードウェアのバージョンを標準出力で表示します。例えば：

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

リファレンスゲノムのロード後、以下のメッセージが標準出力で表示されます。

```
DRAGEN finished normally
```

リファレンスゲノムの準備

DRAGENでリファレンスゲノムを使用できるようにするには、まず、このリファレンスゲノムをFASTA形式からカスタムバイナリー形式に変換しなければなりません。この前処理ステップで使用されるオプションは、性能とマッピングクオリティのバランスをとるためのものです。

DRAGENは、リファレンスゲノムhg19とGRCh37を備えています。どちらのリファレンスゲノムも、一般的な用途に推奨される設定に基づいて、あらかじめインストールされています。これらのリファレンスゲノムを使用することもできますし、リファレンスの前処理コマンドラインオプションを変更することもできます。

ハッシュテーブルのバックグラウンド

DRAGENマッパーは、各リードから重複するシード（サブシーケンスまたはKmer）を抽出します。DRAGENは、PCIeカードのメモリーに存在するハッシュテーブル内のシードを調べ、シードと一致するリファレンスゲノム内の場所を特定します。完全一致を迅速に検索するには、ハッシュテーブルが理想的です。DRAGENハッシュテーブルは、選択されたリファレンスゲノムから構築する必要がありますが、その際に使用されるdragen *--build-hash-table*オプションは、リファレンスゲノムから重複するシードを多数抽出し、ハッシュテーブルのレコードに追加して、このテーブルをバイナリーファイルとして保存します。

リファレンスシードのインターバル

DRAGENハッシュテーブルのサイズは、リファレンスゲノムから追加されたシードの数に比例します。初期設定では、リファレンスゲノムのすべての位置から開始したシードが追加されます。ヒトゲノム1つから約30億個のシードが追加されます。この初期設定では、DRAGEN PCIeボードに32 GB以上のメモリーが必要です。

これよりも大きなヒト以外のゲノムを使用する場合、またはハッシュテーブルの過密度を低減する場合は、*--ht-ref-seed-interval*オプションを使って平均リファレンスインターバルを指定し、追加するリファレンスシードの数を減らします。初期設定で100%追加する場合は、*--ht-ref-seed-interval 1*です。50%の場合は、*--ht-ref-seed-interval 2*を使います。シードのインターバルは整数である必要はありません。例えば、*--ht-ref-seed-interval 1.2*は83.3%追加することを表しますが、この場合、塩基インターバルの平均が1.2になるように、大半は1塩基インターバルで、いくつかの2塩基インターバルが追加されます。

ハッシュテーブルの占有率

ハッシュテーブルにはある一定のサイズが割り当てられますが、必ず多少の空レコードがあるので、占有率は100%を下回ります。DRAGENハッシュテーブルへ迅速にアクセスするには、空のスペースが重要です。レコードは、ハッシュテーブルに疑似ランダムに配置されるため、場所によっては、レコードが異常に密集します。DRAGENは、占有率の上限を90%程度にとどめることを推奨しています。空スペースの割合が0に近づくにつれて、過密領域が大きくなり、DRAGENマッパーによるシードの検索が徐々に遅くなります。

ハッシュテーブル/シード長

ハッシュテーブルには、単一の共通する長さのリファレンスシードが追加されます。この一次シード長は*--ht-seed-len*オプションでコントロールされます。初期設定は21です。

サポートされている一次シード長の最大値は、テーブルのサイズが8～31.5 GBの場合の27塩基です。一般に、シードが長ければ長いほど、ランタイムの性能が、シードが短ければ短いほど、マッピングのクオリティ（成功率と精度）が上がります。長めのシードは、リファレンスゲノムでは一意になることが多いため、代わりの場所をいろいろ確認する必要なく、迅速にマッピングが行われます。しかし、長めのシードは、バリアントやシーケンスエラーなどといった、リファレンスからの逸脱とオーバーラップする可能性が高くなります。これにより、そのシードとの完全一致による正常なマッピングが妨げられますが、それでも、同じリードにある別のシードはマッピング可能で、それぞれのリードで利用可能な長いシード位置は少なくなります。

長めのリードには、逸脱の回避に利用できるシード位置が多いので、長めのシードが適しています。

表 1　シード長の推奨値

*-ht-seed-len*の値	リード長
21	100 bpから150 bp
17から19	より短いリード(36 bp)
27	250 bp超

ハッシュテーブルとシード拡張

反復シーケンスのため、長さに関係なく一部のシードが、リファレンスゲノムの複数の位置に一致することがあります。DRAGENは、シード拡張と呼ばれる独自のメカニズムを使用して、このような高頻度シードをマッピングすることができます。DRAGENが、リファレンスの複数の位置に一次シードが当てはまると判断した場合、DRAGENは、シード長を延長し、リファレンス内で一意になるまで、シードの両端を拡張します。

例えば、21塩基の一次シードの両端を7塩基分延長して、35塩基の拡張シードにすることができます。21塩基の一次シードは、リファレンスの100カ所に一致する可能性があります。これら100カ所のシード位置の35塩基拡張を40グループに分割して、1～3個の同一の35塩基シードにすることができます。DRAGENは、反復シード拡張をサポートします。反復シード拡張は、同じ一次シードのラージセットが、異なる拡張長によって最適に解決される様々なサブセットを含む場合、自動的に行われます。

初期設定では、最大拡張シード長は、一次シード長+128です。拡張シード長の最大値を変更するには、`--ht-max-ext-seed-len`オプションを使用します。例えば、リードが短い場合、DRAGENは、最大拡張シード長をリード長よりも短くすることを推奨します。これは、リード全体よりも長い拡張は決して一致しないからです。

また、以下のオプションを使用して、シードをどの程度、積極的に拡張するかを調整することもできます。これらのオプションは、高度な用途でのみ使用します。

- *--ht-cost-coeff-seed-len*
- *--ht-cost-coeff-seed-freq*
- *--ht-cost-penalty*
- *--ht-cost-penalty-incr*

拡張の長さとヒットの頻度の間にはトレードオフがあります。マッピング速度を向上させるには、長めのシード拡張を使用して、シードのヒット頻度を低減します。また、高いマッピング精度を実現するには、シード拡張を避けるか、短めに抑えながら、ヒット頻度が高くなることを容認します。短めに拡張すると、SNP間でのシードの適合率が高まり、同時にアライメントをスコアリングするためのマッピング位置候補が増えるため、マッピングクオリティが向上します。拡張の初期設定は、シード頻度の初期設定と同様、比較的短いシード拡張と高いヒット頻度でマッピング精度に強く偏った傾向が見られます。

シード頻度オプションの初期設定は以下のとおりです。

オプション	初期設定
--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

シード頻度の上限とターゲット

1つの一次シードまたは拡張シードが、リファレンスゲノムで複数の位置に一致することがあります。このような一致はすべて、ハッシュテーブルに追加され、その後、DRAGENマッパーがリードから抽出された対応するシードを検索するときに取得されます。そのため、アライメントされたマッパー出力を生成するために、複数のリファレンス位置が比較検討されます。しかし、DRAGENでは、シード1つあたりの一致数、つまり頻度が制限されています。この頻度は、`--ht-max-seed-freq`オプションで変更できます。初期設定では、頻度の上限は16です。DRAGENはこれよりも大きな頻度のシードと遭遇した場合、このシードを、特定の拡張シードパターンの頻度が制限内に収まるような十分な長さを持つ二次シードに拡張します。シード拡張を最大にしても上限を超える場合、このシードは却下され、ハッシュテーブルには追加されません。その代わり、DRAGENは高頻度レコードを1つ追加します。

しかし、以下の理由により、このシード頻度の上限が、DRAGENのマッピングクオリティに著しい影響を与えることはあまりありません。

- シードが却下されるのは、シード拡張に失敗した場合のみです。数千カ所に一致する、極端に高頻度の一次シードだけが却下されます。このようなシードはマッピングには使えません。
- 得られたリード内に、確認すべきシード位置がほかにもあります。別のシード位置が1つ以上の一致を返せるほど一意である場合には、このリードを適切にマッピングできます。しかし、すべてのシード位置が高頻度を理由に却下された場合、これは、リード全体が多数のリファレンス位置に一致することを意味する可能性があります。リードがマッピングされたとしても、それはMAPQがゼロかまたは非常に低い、ランダムなマッピングでしょう。

頻度は最大256まで増減できます。頻度の上限を高めに設定すると、マッピングされるリードの数が、特に短いリードの場合、わずかに増える傾向が見られますが、追加でマッピングされたリードのMAPQはゼロかまたは非常に低いものになりがちです。また、これに伴い、多数のマッピング候補が検討されることがあるため、DRAGENマッピングの速度も遅くなる傾向も見られます。

頻度の上限に加え、`--ht-target-seed- freq`オプションを使って、ターゲットのシード頻度を指定することもできます。このターゲット頻度は、高頻度の一次シードの拡張を生成するときに使用されます。拡張の長さは、ターゲットに近い、拡張シード頻度が優先して選択されます。初期設定値は4で、これはDRAGENがシードを一意にマッピングするため、必要以上に短いシード拡張を生成するようにバイアスされていることを意味します。

デコイコンティグへの対応

DRAGENハッシュテーブルビルダーは、ハッシュテーブルを構築する前に、リファレンスからデコイコンティグの欠如を自動的に検出し、このデコイコンティグをFASTAファイルに追加します。デコイファイルは、`/opt/edico/liftover/hs_decoys.fa`にあります。リファレンスにデコイコンティグが存在しない場合、デコイコンティグにマッピングするリードは、出力BAMで人為的にマッピングされていない（unmapped）とマークされます。これは、オリジナルのリファレンスに、デコイコンティグがないからです。その結果、マッピング率が人為的に低くなります。しかし、デコイリードを原因とする誤判断を除去すると、バリアントコーリングの精度が向上します。

初期設定でこの機能を使用することを推奨します。ただし、*--ht-suppress-decoys*オプションをtrueに設定して、これらのデコイがハッシュテーブルに追加されないようにすることができます。

ALTコンティグハッシュテーブル

hg19またはhg38 ALTコンティグが検出されると、ハッシュテーブルビルダーは、リフトオーバーファイルまたはBEDファイルにALTコンティグのマスクを要求します。この要求をオーバーライドするには、ハッシュテーブルを構築するとき、およびDRAGENを実行するときに、*--ht-alt-aware-validate*オプションをfalseに設定します。

ALT-awareハッシュテーブル

DRAGENでALT-awareマッピングを可能にするには、*--ht-alt-liftover*オプションを使用し、リフトオーバーファイルを使ってGRch38（およびALTコンティグを持つその他のリファレンス）を構築します。ハッシュテーブルビルダーは、各リファレンスシーケンスをリフトオーバーファイルに基づいて、primaryとalternateに分類し、alternateの前に、primaryをreference.binにパックします。hg38DHおよびhg19のSAMリフトオーバーファイルは、`/opt/edico/liftover`フォルダーにあります。*--ht-alt-liftover*オプションは、ALT-awareハッシュテーブルを構築するためのリフトオーバーファイルへのパスを指定します。

カスタムリフトオーバーファイル

カスタムリフトオーバーファイルは、DRAGENが提供するリフトオーバーファイルの代わりに使用できます。リフトオーバーファイルはSAM形式でなければなりませんが、SAMヘッダーは必須ではありません。SEQおよびQUALフィールドは省略（*）できます。各アライメントレコードには、QNAMEとして、alternateハプロタイプリファレンスシーケンス名、リファレンスシーケンスにある宛先（通常は、一次アセンブリ）のリフトオーバーアライメントのRNAMEとPOSを示す必要があります。

逆相補方向アライメントは、FLAGの0x10ビットで示されます。マッピングされていない（0x4）または二次（0x100）とフラグ付けされたレコードは無視されます。CIGARには、ハードまたはソフトクリッピングが含まれ、ALTコンティグの一部がアライメントされないまま残されることがあります。

リファレンスシーケンス1つで、ALTコンティグ（QNAMEに現れる）とリフトオーバーの宛先（RNAMEに現れる）の両方の役割を果たすことはできません。複数のALTコンティグを、同一の一次アセンブリ位置にアライメントすることができます。また、1つのALTコンティグに複数のアライメントを提供することもできます（必要に応じて、余剰分にはフラグ0x800が付与されます）。例えば、一部分を順方向、もう一部分を逆相補方向にアライメントできます。ただし、ALTコンティグの各塩基は、その塩基をカバーするM CIGAR操作を持つ最初のアライメントレコードに従って、1つのリフトオーバーイメージしか受け取りません。

SAMレコードのQNAMEがリファレンスゲノムから欠落している場合、このSAMレコードは無視されるので、同じリフトオーバーファイルをさまざまなリファレンスサブセットで使用できます。ただし、QNAMEは持っているが、RNAMEを持たないアライメントがある場合はエラーになります。

ALTをマスクしたハッシュテーブル

ALTコンティグを含む標準的なhg19またはhg38 FASTAからハッシュテーブルを構築するときに、`--ht-alt-liftover`でリフトオーバーファイルを指定しなかった場合、ハッシュテーブルビルダーにより、ALTコンティグの領域が自動的にNでマスクされます。

マスクされた領域の説明は、`/opt/edico/fasta_mask/`にあるBEDファイルに記述されます。また、`hash_table.cfg`には、使用されたBEDファイルのパスと、このファイルの概要が含まれます。`_digest`は、ファイルの特定に使用できる一意の数字を示します。

また、コマンドラインオプション`--ht-mask-bed`を使用して、マスクしたい領域を含むカスタムBEDファイルを設定することもできます。ハッシュテーブルビルダーでは、`--ht-mask-bed`、または`--ht-alt-liftover`のどちらか1つしか使えません。マスクとリフトオーバーファイルの両方を使ってハッシュテーブルを構築するには、`ht-allow-mask-and-liftover`をtrueに設定します。

コマンドラインオプション

リファレンスFASTAをDRAGENマッピングのハッシュテーブルに変換するには、*--build-hash-table*オプションを使用します。このオプションは、連結された複数のリファレンスシーケンスを含むFASTAファイルおよび既存の出力ディレクトリをインプットとして取ります。DRAGENは以下のファイルを生成します。

`reference.bin`	1塩基につき4ビットでコード化されたリファレンスシーケンス。4ビットコードが使用されているため、サイズをバイト単位で表すと、リファレンスゲノムサイズの約半分になります。リファレンスシーケンスのあいだのNはトリミングされ、パディングが自動的に挿入されます。例えば、hg19には、93シーケンスに3137161264個の塩基があります。これは1526285312バイト= 1.46 GBにコード化されます。ここで、1 GBは1ギガバイト、または2^{30}バイトを意味します。
`hash_table.cmp`	圧縮されたハッシュテーブル。ハッシュテーブルは、DRAGENマッパーによって展開され、`--ht-seed-len`オプションで指定された長さを持つ一次シードとさまざまな長さを持つ拡張シードの検索に使用されます。
`hash_table.cfg`	生成されたハッシュテーブルのパラメーターと属性のリスト（テキスト形式）。このファイルには、リファレンスゲノムとハッシュテーブルに関する主要な情報が記載されています。
`hash_table.cfg.bin`	`hash_table.cfg`のバイナリー版。DRAGENの構成に使用されます。
`hash_table_stats.txt`	ハッシュテーブルの占有率を含め、構成されたハッシュに関する詳細な内部統計を列挙したテキストファイル。このテーブルは情報提供のみを目的とし、ほかのツールでは使用されません。

Buildコマンドの使用法は以下のとおりです。

```
dragen --build-hash-table true [options] --ht-reference <reference.fasta> --output-directory <outdir>
```

これ以降のセクションでは、ハッシュテーブルの構築に使用されるオプションについて説明します。

インプット/アウトプットオプション

ハッシュテーブルの構築には、`--ht-reference`と`--output-directory`オプションが必要です。`--ht-reference`オプションには、リファレンスFASTAファイルへのパスを、`--output-directory`には、ハッシュテーブル出力ファイルの書き込み先となる既存のディレクトリを指定します。構築した個々のハッシュテーブルは、別々のフォルダーに整理することを推奨します。ハッシュテーブルを生成したときに初期設定以外のパラメーター設定を使用した場合、それをフォルダー名に付け加えるのがベストプラクティスです。リファレンスFASTAファイルのシーケンス名は一意でなければなりません。

一次シード長

`--ht-seed-len`オプションは、リファレンスゲノムからハッシュテーブルに追加するシードのヌクレオチドの初期長を指定します。ハッシュテーブルでシードの編集が有効化されている場合を除き、マッパーはランタイムに、各リードから同じ長さのシードを抽出し、完全一致を検索します。

一次シード長の最大値は、ハッシュテーブルサイズの関数です。テーブルサイズが16～64 GBの場合、上限はk=27であり、これはヒト全ゲノムの典型的なサイズに対応します。また、テーブルサイズが4～16 GBの場合の上限はk=26です。

一次シード長の最小値は、主にリファレンスゲノムのサイズと複雑さによって決まります。シード長は、大半のリファレンス位置を一意に解決できる十分な長さを必要とします。ヒト全ゲノムのリファレンスでは、ハッシュテーブルの構築は通常、k < 16で失敗します。ゲノムが短めの場合は下限を小さく、ゲノムがあまり複雑ではない（反復性が高い）場合は大きくします。log4(3.1 G) ≈ 16なので、3.1 Gbpのヒトゲノムに対する一意性の閾値が`--ht-seed-len 16`であることは直感的に理解できます。つまり、3.1 Gのリファレンス位置を区別するには、4つのヌクレオチドから少なくとも16個の選択が必要です。

精度についての注意点

リードのマッピングを正常に行うには、少なくとも1つの一次シードが完全一致、また、編集済みシードを使用している場合には単一のSNPと一致していなければなりません。短めのシードは、リファレンスへのマッピングが成功しやすい傾向にあります。これは、各リードに適合するシードの数が増え、シードがバリアントと重複したり、シーケンシングエラーとなったりする可能性が低くなるからです。

ただし、シードが非常に短いと、マッピング精度が低下することもあります。非常に短いシードは複数のリファレンス位置にマッピングされることがよくあり、マッパーがマッピング位置エラーの増加とみなす原因となります。Smith-Watermanアライメントスコアリングなどのヒューリスティックが変異やエラーを不完全にモデリングするため、このように誤った一致がレポートされることがあります。`-- Aligner.aln_min_score`のようなランタイムのクオリティフィルターは、非常に短いシードを使って、精度の問題をコントロールできます。

スピードについての注意点

短めのシードがマッピング速度を低下させることはよくあります。これは、シードがマッピングされるリファレンス位置が増えるため、最適な結果を判断するためにSmith-Watermanアライメントを行うなど、追加の作業が発生するからです。この影響は、一次シード長がリファレンスゲノムの一意性の閾値（例えば、ヒト全ゲノムの場合K=16）に近づいたときに、最も顕著に表れます。

利用時の注意点

- **リード長**：一般に、短めのリードには短めのシードが、長めのリードには長めのシードが適しています。短いリードの中に、バリアントやシーケンスエラーによるミスマッチがあると、リードが参照と一致する短いセグメントだけに切り分けられることがあります。これにより、短いシードのみがその差分にフィットし、リファレンスと正確に一致することができます。例えば、36 bpリードでは、中央にある1個のSNPが18 bpを超える長さのシードをブロックし、リファレンスとの照合を妨げます。250 bpのリードで、27 bpシードをブロックする確率が0.01%を超すには、SNPが15個必要です。
- **ペアエンド**：ペアエンドリードを使用すると、長めのリードでのマッピング精度が向上します。DRAGENは、マッピング精度の向上にペアエンド情報を使用します。これには、ある特定のリファレンス領域に対するシードマッピングを持っているのが1つのメイトだけである場合に、予想されるリファレンスウィンドウを検索するレスキュースキャンの使用が含まれます。したがって、ペアエンドリードでは、完全一致シードが正確なアライメントを見つける確率が2倍になります。
- **バリアントまたはエラー率**：リファレンスとリードの差が頻繁に発生する場合、指定されたリードの差分位置の間に収まり、リファレンスと完全一致する短めのシードが必要になります。
- **マッピング率の要件**：MAPQが低くても、高いパーセンテージのリードをマッピングする必要がある事例では、短いシードが役に立つでしょう。リファレンスとうまく一致しないリードの中には、短いシードを使って、リファレンスとの部分一致を見つけるとマッピングするものもあります。

最大シード長

*--ht-max-ext-seed-len*オプションは、ハッシュテーブルに追加される拡張シードの長さを制限します。多数のリファレンス位置に一致する一次シードを拡張して、一致の一意性を高めることができますが、そのためには最大ヒット頻度（*--ht-max-seed-freq*）内でシードをマッピングする必要があります。一次シードの長さは、*--ht-seed-len*を使って指定します。

一次シード長がkの場合、最大シード長は、kからk+128の間で指定します。初期設定は上限のk+128です。

シード拡張の制限

50 bp未満のリードなど、短いリードについては、*--ht-max-ext-seed-len*オプションを推奨します。短いリードの場合、シード拡張をリード長よりも1– 4 bpのように、わずかに短く制限するといいでしょう。例えば、36 bpリードの場合、*--ht-max-ext-seed-len*を35に設定します。この設定により、ハッシュテーブルビルダーがリードより長いシード拡張を計画します。長いシード拡張は、ラン中に、短い拡張でリードに収まったかもしれないシードの、シード拡張とマッピングの失敗を引き起こす可能性があります。

同様に、長めのリードのシード拡張も制限できます。例えば、100 bpのリードに対し、*--ht-max-ext-seed-len*を99に設定します。シードは何があっても控えめに拡張されるので、長めのリードのシード拡張を制限しても、それほどの有効性は得られません。初期設定の上限値k+128であっても、個々のシードは、最大ヒット頻度（*--ht-max-seed-freq*）を下回るように拡張されますし、上回る場合でも拡張は数塩基分に抑えられます。これは、ターゲットのヒット頻度（*--ht-target-seed-freq*）に近づけるため、または段階的な拡張ステップが多くなりすぎるのを避けるためです。

最大ヒット頻度

*--ht-max-seed-freq*オプションは、シードのヒット数（リファレンスゲノムの位置）に対する制限を設定します。この制限は、どのような一次または拡張シードに対しても追加できます。ある一次シードが、指定された制限よりも多くのリファレンス位置にマッピングされた場合、拡張シードが、制限を下回る、小さな同一のシードグループに分割されるように、この一次シードを十分に長く拡張しなければなりません。最大拡張シード長（*--ht-max-ext-seed-len*）であっても、同一のリファレンスシードのグループがこの制限値よりも大きい場合、それらのリファレンス位置は、ハッシュテーブルに追加されません。その代わり、DRAGENは高頻度レコードを1つ追加します。

最大ヒット頻度は、1から256の間で設定します。この値が小さすぎると、必要となるシード拡張の数が多すぎて、ハッシュテーブルの構築がエラーとなります。ヒト全ゲノムリファレンスの最小推奨値は8です。

精度についての注意点

最大ヒット頻度が多ければ多いほど、マッピングは成功します。

- 制限値が高ければ高いほど、その制限下でマッピングできないリファレンス位置が少なくなります。
- 制限値が高ければ高いほど、シード拡張が短くなります。これにより、バリアントまたはシーケンスエラーと重複しない、完全なシード一致の確率が高まります。

ただし、非常に短いシードと同様、ヒット数を大きくすると、マッピング精度が低下することもあります。大きなグループにあるシードヒットの大半は、本当のマッピング位置ではありません。不完全なスコアリングモデルが原因で、ときどき、このような誤ったヒットがレポートされることがあります。また、マッパーも、検討の対象となるリファレンス位置の総数を制限します。あまりにも多数のヒット数を許容すると、実際の最高の一致が検討から締め出される可能性があります。

スピードについての注意点

最大ヒット頻度を高めに設定すると、リードのマッピング速度が低下することはよくあります。これは、シードマッピングで発見されるリファレンス位置が増えるため、最適な結果を判断するためにSmith-Watermanアライメントを行うなど、追加の作業が発生するからです。

ALTコンティグハッシュテーブルのオプション

ここでは、以下のALTコンティグハッシュテーブルオプションについて説明します。ALTコンティグを含むリファレンスからのハッシュテーブルの構築について、詳しくは、14 ページの 「ALTコンティグハッシュテーブル」 を参照してください。

- *--ht-alt-liftover*

  --ht-alt-liftover オプションは、ALT-aware ハッシュテーブルを構築するためのリフトオーバーファイルへのパスを指定します。ALT コンティグを持つリファレンスから構築する場合は、このオプションが必要です。hg38DH および hg19 の SAM リフトオーバーファイルは、`/opt/edico/liftover` フォルダーにあります。

- *--ht-alt-aware-validate*

  ALT コンティグを含むリファレンスからハッシュテーブルを構築するときには、リフトオーバーファイルを使う必要があります。この要件を無効化するには、*--ht-alt-aware-validate* オプションを false に設定します。

- *--ht-decoys*

  DRAGEN は、hg19 および hg38 リファレンスの使用を自動的に検出し、FASTA ファイルにデコイがなければ、ハッシュテーブルに追加します。デコイファイルへのパスを指定するには、*--ht-decoys* オプションを使用します。初期設定は `/opt/edico/liftover/hs_decoys.fa` です。

- *--ht-suppress-decoys*

  ハッシュテーブルを構築するときに、デコイファイルの使用を抑えるには、*--ht-suppress-decoys* オプションを使用します。

- *ht-mask-bed*

  マスクする ALT コンティグ領域を含むカスタム BED ファイルを指定するには、*ht-mask-bed* オプションを使用します。

- *ht-allow-mask-and-liftover*

  マスクとリフトオーバーファイルの両方を使ってハッシュテーブルを構築するには、*ht-allow-mask-and-liftover* オプションを使用します。

DRAGENのソフトウェアオプション

- *--ht-num-threads*

  --ht-num-threads オプションは、ハッシュテーブルの構築を加速するために使用されるワーカー CPU スレッド数の最大値を決定します。このオプションの初期設定は 8 で、最大 32 スレッドまで指定できます。

使用しているサーバーでこれ以上のスレッドの実行がサポートされている場合、最大値を使用することを推奨します。例えば、DRAGEN サーバーはハイパースレッドが可能な 24 コアを搭載していますから、値には 32 を使用すべきです。大きな値を使用するときには、*--ht-max-table-chunks* も調整する必要があります。このサーバーでは 128 GB のメモリーを使用できます。

- *--ht-max-table-chunks*

  --ht-max-table-chunks オプションは、メモリー内に同時に存在する約 1 GB のハッシュテーブルのチャンク数を制限することにより、ハッシュテーブル構築中のメモリーフットプリントをコントロールします。構築中、チャンクが 1 つ増えるたびに、そのおよそ 2 倍（約 2 GB）のシステムメモリーが消費されます。

  ハッシュテーブルは、2 の累乗個の独立したチャンクに分割されます。チャンクのサイズ X は固定で 0.5 GB < X ≤ 1 GB ですが、ハッシュテーブルのサイズによって異なります。例えば、24 GB のハッシュテーブルには、0.75 GB の独立したチャンクが 32 個含まれますが、これは十分なメモリを持つ並列スレッドで構築されます。また、16 GB のハッシュテーブルには、1 GB の独立したチャンクが 16 個含まれます。

  初期設定は *--ht-max-table-chunks* で、これは *--ht-num-threads* と同じですが、*--ht-max-table-chunks* の初期設定の最小値は 8 です。ハッシュテーブルのチャンクを 1 つ作成するには、メモリーにチャンクスペースが 1 つと、それに作用するスレッドが 1 つ必要であるため、初期設定では、これらのオプションは一致しています。しかし、*--ht-max-table-chunks* を *--ht-num-threads* よりも、または *--ht-num-threads* を *--ht-max-table-chunks* よりも大きくすると、構築速度の面で有利です。

サイズオプション

- *--ht-mem-limit*：メモリーの上限

  --ht-mem-limit オプションは、ハッシュテーブルとコード化されたリファレンスゲノムの両方で使用可能な DRAGEN ボードメモリーを指定することにより、生成されたハッシュテーブルのサイズをコントロールします。*--ht-mem-limit* オプションの初期設定は、リファレンスゲノムが WHG のサイズに近い場合は 32 GB、リファレンスのサイズがそれよりも小さい場合は十分な余裕のあるサイズとなります。通常、これらの初期設定をオーバーライドする理由はほとんどありません。

- *--ht-size*：ハッシュテーブルサイズ

  リファレンスゲノムのサイズと使用可能なメモリー（*--ht-mem-limit*）から適切なハッシュテーブルのサイズを計算するのではなく、指定する場合は、このオプションを使用します。テーブルサイズの決定には、初期設定の方法を使用することが推奨されます。*--ht-mem-limit* は二次的な方法として使用してください。

シード追加オプション

- *--ht-ref-seed-interval*：シードインターバル

  --ht-ref-seed-interval オプションは、ハッシュテーブルに追加されたリファレンスゲノムのシードの位置と位置の間隔を定義します。インターバル 1（初期設定）は、すべてのシード位置にシードが追加されることを、また、2 は位置の 50% に追加されることを表します。非整数値もサポートされています。例えば、2.5 を指定すると、位置の 40% にシードが追加されます。

  ヒトの全リファレンスからのシードを 100% 追加するには、DRAGEN ボードに 32 GB のメモリーが必要です。かなり大きなリファレンスゲノムを使用する場合は、サイズに応じて、このオプションを変更します。

- *--ht-soft-seed-freq-capと--ht-max-dec-factor：*シードを間引きするためのソフト頻度上限と最大間引き係数

  シードの間引きは、高頻度の領域でマッピング性能を向上させるための実験的手法です。一次シードの頻度が、*--ht-soft-seed- freq-cap* オプションで指定された上限を超える場合、上限を上回らないように、シード位置の一部だけにシードが追加されます。*--ht- max-dec-factor* オプションは、シードの最大間引き係数を表します。例えば、*--ht-max-dec-factor 3* では、オリジナルのシードの 1/3 以上が保持されます。また、*--ht-max-dec-factor 1* では間引きは行われません。

  間引きは、シードとシードの間に大きなギャップが残らないよう、慎重に行われます。シードの間引きにより、間引きなしにはヒット頻度の上限を超えてしまうような高頻度リファレンス領域でも、シードカバレッジのマッピングを実現できます。また、シードの間引きにより、シード拡張を短く保つことができるため、マッピングの成功率が上昇します。現在までのテストでは、シードの間引きが、その他の精度最適化メソッドよりも優れているということは証明されていません。

- *--ht-rand-hit-hifreqと--ht-rand-hit-extend：*HIFREQレコードとEXTENDレコードを使ったランダムサンプルヒット

  HIFREQ または EXTEND レコードがハッシュテーブルに追加されたときには必ず、特定のシードに対するリファレンスヒットの大きなセットの代わりに、このレコードが有効になります。必要に応じて、ハッシュテーブルビルダーは、そのセットからランダムな代表値を選択し、その HIT レコードを、HIFREQ または EXTEND レコードとともに追加することもできます。

  ランダムサンプルヒットの提供する代替アライメントは、レポートされたアライメントの MAPQ を正確に推測するのに役立ちます。ランダムサンプルヒットが、これ以外の状況で、アライメント位置のレポートに使用されることはありません。これは、ランダムサンプルヒットが結果として、ハッシュテーブルの構築中に選択された位置のバイアスカバレッジになるからです。

  サンプルヒットを含めるには、*--ht-rand-hit-hifreq* を 1 に設定します。*--ht-rand-hit-extend* オプションは、サンプルヒットに含める拡張前ヒット数の最小値です。0 の場合、このオプションは無効化されます。これらのオプションの変更は推奨されません。

シード拡張コントロール

DRAGENシード拡張は動的で、リファレンス位置のマッピングが多すぎるKmerに対し、必要に応じて適用されます。シードは、一次シード長から最大長まで、2～14塩基ずつ（必ず偶数）、段階的に拡張されます。塩基は拡張ステップごとに対称的に付加されます。次の拡張がある場合は、これによりその増分が決まります。

高頻度一次シードそれぞれに、複雑なシード拡張ツリーが関連付けられていることがあります。各フルツリーはハッシュテーブルの構築中に生成され、ルートからのパスは、シードのマッピング中に、反復拡張ステップによりトレースされます。ハッシュテーブルビルダーは、動的なプログラミングアルゴリズムを使用して、可能性のあるあらゆるシード拡張ツリーのスペースから最適なものを探します。このとき、マッピング精度と速度のバランスをとるコスト関数が使用されます。コスト関数を定義するオプションは以下のとおりです：

- *--ht-target-seed-freq：*ターゲットのヒット頻度

*--ht-target-seed-freq*オプションは、シード拡張が目標とすべき、1シードあたりの理想的なヒット数を定義します。値が大きければ大きいほど、最終的なシード拡張が少なく、短くなります。これは、シードが短ければ短いほど、一致するリファレンス位置の数が多くなる傾向にあるからです。

- *--ht-cost-coeff-seed-len*：シード長のコスト係数

  --ht-cost-coeff-seed-len オプションは、シード拡張の基準となる塩基ごとにコストコンポーネントを割り当てます。追加される塩基はコストとみなされます。これは、シードが長くなればなるほど、バリアントの重複やシーケンスエラーが発生し、正しくマッピングされないリスクが高まるからです。値が大きければ大きいほど、最終的なシード拡張が短くなります。

- *--ht-cost-coeff-seed-freq*：ヒット頻度のコスト係数

  --ht-cost-coeff-seed-freq オプションは、ターゲットのヒット頻度と 1 つのシードに対して追加されるヒット数の差にコストコンポーネントを割り当てます。値が大きければ大きいほど、高頻度シードがさらに拡張され、頻度はターゲットに向けて小さくなる傾向がよく見られます。

- *--ht-cost-penalty*：シード拡張のコストペナルティ

  --ht-cost-penalty オプションは、一次シード長を超える拡張に対し、一律にコストを割り当てます。大きな値を設定すると、拡張されるシードの数が減ります。初期設定値は 0 です。

- *--ht-cost-penalty-incr*：拡張ステップのコスト増分

  --ht-cost-penalty-incr オプションは、一次シード長から拡張シード長までにかかった段階的なシード拡張ステップそれぞれに対して、反復コストを割り当てます。ステップ数が多いほど、考慮されるコストが高くなります。これは、小さなステップを多数使って拡張すると、EXTEND 中間レコードが必要とするハッシュテーブルスペースが増えるため、拡張の実行にかかるランタイムも大幅に長くなるからです。値を大きくすれば、シード拡張ツリーのノードが少なくなり、ルートの一次シード長からリーフの拡張シード長まで、より長いステップで、ステップ数を少なくして到達できるようになります。

パイプライン固有のハッシュテーブル

RNA-Seq

初期設定では、DRAGENは、ハッシュテーブルを構築するときにDNA解析のオプションを設定します。RNA-Seqデータを実行するには、`--ht-build-rna-hashtable`をtrueに設定して、RNA-Seqハッシュテーブルを構築する必要があります。RNA-Seqアライメントを実行する場合、自動生成されるサブディレクトリではなく、オリジナルの`--output-directory`を使用します。

CNV

CNVパイプラインを使用する場合、`--enable-cnv`をtrueに設定します。このコマンドは、CNVアルゴリズムで使用される追加のKmerハッシュマップを生成します。必ず、`--enable-cnv`オプションを使用することを推奨します。これにより、マッピングとアライメントで同じハッシュテーブルを使って、CNVコールを実行できます。

メチル化

メチル化パイプラインを実行するには、メチル化特異的なハッシュテーブルを構築する必要があります。DRAGENは、シングルパス、または従来のマルチパスのメチル化ハッシュテーブルを構築できます。メチル化の実行は、シングルパスのハッシュテーブルを使用するほうが従来のマルチパスのハッシュテーブルよりも速く完了します。メチル化テーブルの構築と解析の実行には、シングルパスのハッシュテーブルの使用を推奨します。

ハッシュテーブルのタイプ	ハッシュテーブルのコマンド
シングルパス	• --ht-methylated-combined=true • --ht-seed-len 27
マルチパス	• --ht-methylated=true • --ht-seed-len 27 • --ht-max-seed-freq 16

シングルパスのハッシュテーブル

以下に、シングルパスのハッシュテーブルを構築する例を示します。この例では、リファレンスインデックスフォルダーの`methyl_converted`サブディレクトリの下に統合ハッシュテーブルが生成されます。

```
dragen --build-hash-table true \ --output-directory $REFDIR \ --ht-reference $FASTA \ --ht-num-threads 40 \ --ht-methylated-combined=true \ --ht-seed-len 27
```

マルチパスのハッシュテーブル

メチル化を実行するには、CをTに変換したリファレンス塩基を持つハッシュテーブルと、GをAに変換したリファレンス塩基を持つハッシュテーブルの2つを構築する必要があります。`--ht-methylated`コマンドラインオプションを使用すると、この変換が自動的に行われます。変換後のハッシュテーブルは、`--output-directory`コマンドラインオプションで指定されたフォルダーの下にある2つのサブディレクトリに生成されます。これらのサブディレクトリには、塩基の変換に合わせて`CT_converted`と`GA_converted`という名前が付けられます。ハッシュテーブルを使用して、メチル化アライメントを実行する場合には、必ず、サブディレクトリではなく、`--output-directory`フォルダーを参照します。

塩基変換により、ハッシュテーブルから相当な量の情報が削除されます。従来のハッシュテーブルの構築とは異なるハッシュテーブルパラメーターの使用が必要になる場合もあります。哺乳類種のハッシュテーブルを構築する場合は、以下のオプションの使用を推奨します。

```
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
```

入力ファイルと出力ファイルの場所の決定

DRAGEN Bio-It Platformは極めて高速なので、入力ファイルと出力ファイルの場所は慎重に計画する必要があります。インプットまたは出力ファイルが保管されているファイルシステムが低速であった場合、システム全体のパフォーマンスが、そのファイルシステムのスループットによって制限されてしまいます。外部ストレージシステムをマウントし、このシステムとの間で直接、インプットとアウトプットをストリームすることを推奨します。

パフォーマンスを確保するため、DRAGENシステムには、RAID-0でグループ化された高速SSDディスクセットで構成される高速ファイルシステムがあらかじめ1つ以上設定されています。このファイルシステムは、`/staging`にマウントされます。この領域は大きく高速ですが、冗長性がないことを強調するために、この名前がつけられました。このファイルシステムを構成するディスクのいずれかで不具合が発生すると、保存されているデータがすべて失われます。

処理の間、DRAGENは、一時ファイルを生成し、リードバックします。DRAGENでは、*--intermediate-results-dir*オプションを使用して、一時ファイルを必ず高速SSD（または`/staging`）へ移動することを強く推奨します。*--intermediate-results-dir*オプションが指定されていない場合、一時ファイルは`--output-directory`に書き込まれます。外部ストレージシステムをマウントし、このシステムとの間でインプットとアウトプットをストリームすることを推奨します。

インプットデータの処理

FASTQデータを解析するには、dragenコマンドを使用します。例えば、シングルエンドのFASTQファイルを解析するには、次のコマンドを使用します：

```
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
```

コマンドラインオプションの詳細については、48 ページの 「DRAGENホストソフトウェア」 を参照してください。

FASTQデータ処理のためのコマンド例

リファレンスをロードしたら、インプットFASTQデータを処理できます。データセットに最適な例を選択してください。30xカバレッジでヒト全ゲノムをエンドツーエンドで（FASTQインプットからVCFアウトプットまで）処理するために、SSDドライブ付きの24コアサーバーで、これらのコマンドを実行したときの所要時間は最大約30分です。速度は、インプットのサイズに左右されるので、例えば、60xカバレッジは2倍の時間がかかります。エクソームデータはあまり時間がかかりません。正常終了すると、以下のメッセージ（スクリプトから実行した場合はアプリケーションの終了コード0）が表示されます：

```
DRAGEN finished normally
```

このメッセージに続けて、リードカウントやパフォーマンスなどのメトリクスがまとめて表示されます。コマンドラインオプションに問題があった場合、エラーに続けて、使い方のヘルプが表示されます。エラーの確認には、スクロールが必要になることがあります。

DRAGENのログをファイルにリダイレクトし、今後の参考として保存することができます。

dragenコマンドラインオプションのヘルプを表示するには、以下のコマンドを実行します：

```
dragen -h
```

本文書に記載されたコマンド例は、読みやすいように書式設定されているため、改行文字が含まれています。コピーアンドペーストによるエラーを回避するために、ここで説明するコマンド例は、`/opt/edico/examples/`のシェルスクリプトに保存されています。これらのコマンド例の要件は以下のとおりです：

- すべてのコマンドは、FASTQ、またはgzipされたFASTQ（fastq.gz）に対応します。ファイルの種類は、DRAGENにより自動的に判断されます。
- すべてのコマンドに、*-f*オプションが含まれます。つまり、すでに出力ファイルが存在する場合は、強制的に上書きされます。
- すべてのコマンドは、DRAGENリファレンス（ハッシュテーブル）ディレクトリが/staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149、FASTAリファレンスファイルが/staging/human/reference/hg19/hg19.faであることを前提にしています。必要に応じて、これらを、有効なリファレンスまたはディレクトリパスで置き換えることができます。
- コマンド例はすべて、サンプルデータパッケージが`/staging/examples`に保管されていることを前提にしています（特に.fastqファイルとfastq.gzファイルは`/staging/examples/reads`に保管されていることが前提です）。
- これらのコマンド例を実行するには、`/staging/examples`フォルダーへの書き込みアクセス権が必要です。

エンドツーエンドのアライメントとバリアントコーリングの例

ペアエンドBAMインプット、VCFアウトプット

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-b /staging/human/unsorted_SRA056922_30x_e10_50M.bam \
--enable-map-align true \
--enable-map-align-output true \
--enable-variant-caller true \
--vc-sample-name Unsorted_SRA056922_30x_e10_50M \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M \
--enable-duplicate-marking true
```

- または、`/opt/edico/examples/paired_fastq_in_dupmark_bam_and_vcf_out.sh`を実行します。

上の例で、*/staging/human/unsorted_SRA056922_30x_e10_50M.bam*入力ファイルが存在しない場合は、`/opt/edico/examples/paired_fastq_in_unsorted_bam_out.sh`スクリプトを実行して、生成します。

ペアエンドFASTQインプット、VCFアウトプット（初期設定）

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
```

```
--enable-variant-caller true \
--RGID Illumina_RGID \
--RGSM SRA056922_30x_e10_50M \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M \
```

- または、/opt/edico/examples/paired_fastq_in_vcf_out.shを実行します。

この例は、エンドツーエンドで実行するために指定しなければならない最小限のオプションを示しています。初期設定では、重複マーキングは実行されず、BAMアウトプットは作成されません。

ペアエンドFastqインプット、ソート済みで重複マーキング済みのVCFアウトプット

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--enable-variant-caller true \
--RGID Illumina_RGID \
--RGSM SRA056922_30x_e10_50M \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M \
--enable-duplicate-marking true
```

- または、/opt/edico/examples/paired_fastq_in_dupmark_vcf_out.shを実行します。

ペアエンドFASTQインプット、ソート済みBAMとVCFアウトプット

- 以下のとおり入力します：

```
dragen -f
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--enable-variant-caller true \
--RGID Illumina_RGID \
--RGSM SRA056922_30x_e10_50M \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M \
--enable-duplicate-marking true \
--enable-map-align-output true
```

- または、/opt/edico/examples/sorted_bam_in_vcf_out.shを実行します。

ペアエンドFASTQインプット、ソート済みSAMとVCFアウトプット

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--enable-variant-caller true \
--RGID Illumina_RGID \
--RGSM SRA056922_30x_e10_50M \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M \
--enable-duplicate-marking true \
--enable-map-align-output true \
--output-format SAM
```

- または、/opt/edico/examples/paired_fastq_in_dupmark_sam_and_vcf_out.shを実行します。

ペアエンドFASTQインプット、ソート済みCRAMとVCFアウトプット

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--enable-variant-caller true \
--RGID Illumina_RGID \
--RGSM SRA056922_30x_e10_50M \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M \
--enable-duplicate-marking true \
--enable-map-align-output true \
--output-format CRAM \
```

- または、/opt/edico/examples/paired_fastq_in_dupmark_cram_and_vcf_out.shを実行します。

ペアエンドFASTQインプット、ソート済みBAMとVCFアウトプット、リピートジェノタイピングVCF

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--enable-variant-caller true \
--RGID Illumina_RGID \
```

```
--RGSM SRA056922_30x_e10_50M \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M \
--enable-duplicate-marking true \
--enable-map-align-output true \
--repeat-genotype-enable true \
--repeat-genotype-specs /opt/edico/repeat-specs/hg19 \
--repeat-genotype-sex female \
--repeat-genotype-ref-fasta /staging/human/reference/h19/hg19.fa
```

アライメントのみの例

これらの例に示されたアライメントのバリエーションはすべて、エンドツーエンドのケースでも使用できます。

マッピング/アライメント：シングルエンドのFASTQインプット、ソート済みBAMアウトプット（初期設定）

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_rand1_100K.fastq \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_rand1_100K \
--RGID DRAGEN_RGID \
--RGSM DRAGEN_RGSM \
```

- または、`/opt/edico/examples/single_fastq_in_bam_out.sh`を実行します。

マッピング/アライメント：シングルエンドのFASTQインプット、ソート済みで重複マーキング済みのBAMアウトプット

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_rand1_100K.fastq \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_rand1_100K_dup_marked \
--enable-duplicate-marking true \
--RGID DRAGEN_RGID \
--RGSM DRAGEN_RGSM
```

- または、`/opt/edico/examples/single_fastq_in_dupmark_bam_out.sh`を実行します。

マッピング/アライメント：ペアエンドのFASTQインプット、ソート済みBAMアウトプット（初期設定）

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M \
--RGID DRAGEN_RGID \
--RGSM DRAGEN_RGSM
```

- または、/opt/edico/examples/paired_fastq_in_bam_out.shを実行します。

マッピング/アライメント：ペアエンドのFASTQインプット、ソート済みCRAMアウトプット

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M \
--output-format CRAM \
--RGID DRAGEN_RGID \
--RGSM DRAGEN_RGSM
```

- または、/opt/edico/examples/paired_fastq_in_cram_out.shを実行します。

マッピング/アライメント：ペアエンドのFASTQインプット、ソート済みで非圧縮のBAMアウトプット

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--output-directory /staging/examples/ \
--output-file-prefix uncompressed_SRA \
--enable-bam-compression false \
--RGID DRAGEN_RGID \
--RGSM DRAGEN_RGSM
```

マッピング/アライメント：ペアエンドのFASTQインプット、ソート済みSAMアウトプット

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M \
--output-format SAM \
--RGID DRAGEN_RGID \
--RGSM DRAGEN_RGSM
```

- または、/opt/edico/examples/paired_fastq_in_sam_out.shを実行します。

アライメント：ペアエンドのFASTQインプット、未ソートのBAMアウトプット

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--output-directory /staging/examples/ \
--output-file-prefix unsorted_SRA056922_30x_e10_50M \
--enable-sort false \
--RGID DRAGEN_RGID \
--RGSM DRAGEN_RGSM
```

- または、/opt/edico/examples/paired_fastq_in_unsorted_bam_out.shを実行します。

マッピング/アライメント：マージしたペアエンドのFASTQインプット、

BAMアウトプット

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_PE_30x_rand1_10K_interleaved.fastq \
--interleaved \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_PE_30x_rand1_10K_interleaved \
--RGID DRAGEN_RGID \
--RGSM DRAGEN_RGSM
```

- または、/opt/edico/examples/interleaved_fastq_in_bam_out.shを実行します。

RNAマッピングとアライメントのみの例

マッピング/アライメントのみの例はすべて、RNAに使用できます。コマンドで違う点は、*--enable-rna*オプションをtrueに設定するところだけです。DRAGENは、自動的にRNA専用のハッシュテーブルを選択し、RNAスプライスアライナーを使用して処理します。

これらの例で使用されるハッシュテーブルは、*--ht-build-rna-hashtable true*オプションを使って生成する必要があります。そうしないと、実行が失敗し、次のようなエラーが表示されます。

```
ERROR:The specified hashtable directory cannot be used to run RNA:
/staging/examples/reference/hg19/hg19.fa.k_21.f_16.m_149
```

このようなエラーが表示された場合、*--ht-build-rna-hashtable true*オプションを使って、ハッシュテーブルを再度生成します。

RNAマッピング/アライメント：ペアエンドのFASTQインプット、BAMアウトプット

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M \
--enable-rna true \
--RGID DRAGEN_RGID \
--RGSM DRAGEN_RGSM
```

以下のコマンドライン例では、遺伝子アノテーションの入ったgtfファイルへのパスを含む追加コマンドラインオプションを使って、RNA-Seqデータがマッピング/アライメントされます。遺伝子アノテーションファイルにより、（新たに検索するのではなく）既知のスプライスジャンクションの一覧が提供されるため、マッピングが強化されます。

```
dragen -f \
-r <HASHTABLE_DIR>
-1 <FASTQ1> \
-2 <FASTQ2> \
-a $/reference_genomes/annotation/GTF/$gencode.annotation.gtf
--enable-map-align true \
--enable-sort=true \
--enable-bam-indexing true \
--enable-map-align-output true \
--output-format=BAM \
--RGID=<READ_GROUP_ID> \
--RGSM=<Sample_NAME> \
--RGPL=<LIBRARY> \
--config-file /opt/edico/config/dragen-user-defaults.cfg \
--enable-rna=true \
--output-directory <OUT_DIR> \
--output-file-prefix <PREFIX>
```

RNA定量

遺伝子と転写産物の発現定量を実行するには、次のオプションを追加します：

```
--enable-rna-quantification true
```

*--enable-rna-quantification*がtrueに設定されている場合、GCバイアス補正が初期設定で、*--rna-quantification-gc-bias*を有効化する必要はありません。また、ライブラリータイプは自動的に検出されるので、*--rna-quantification-library-type*を設定する必要はありません。

ご注意

ライブラリータイプの自動検出は、ペアエンドデータでのみ機能します。シングルエンドデータの場合、*--rna-quantification-library-type*オプションを設定して、ライブラリーを指定する必要があります。

RNA Fusion

遺伝子融合検出を実行するには、次のオプションを追加します：

```
--enable-rna-gene-fusion true
```

融合ではライブラリーは使用されないので、指定する必要はありません。

エピゲノムマッピングとアライメントの例

バイサルファイトシーケンシングデータを使ったエピゲノム（メチル化）マッピングとアライメントを実行する前に、まず、次のようにして、メチル化専用のリファレンスハッシュテーブルを作成する必要があります：

```
mkdir -p /staging/human/reference/hg19_epigenome
dragen --build-hash-table true \
--ht-reference /staging/human/reference/hg19/hg19.fa \
--ht-max-seed-freq 64 --ht-seed-len 27 --ht-methylated true \
--output-directory /staging/human/reference/hg19_epigenome \
--ht-alt-liftover /opt/edico/liftover/hg19_alt_liftover.sam
```

前述のdragenコマンドは、`/staging/human/reference/hg19_epigenome`の下に、`GA_converted`と`CT_converted`の2つのハッシュテーブルディレクトリを作成します。`CT_converted`ハッシュテーブルは、リファレンスシーケンスの各C塩基をTに変換することにより作成されます。同様に、`GA_converted`ハッシュテーブルは、リファレンスシーケンスの各G塩基をAに変換することにより作成されます。塩基変換により作成されたリファレンスは複雑さが緩和されます。また、埋め合わせのため、哺乳類ゲノムでは通常、ハッシュテーブルのシード長引数（*--ht-seed-len*）が27に増やされます（シード長の初期設定は21）。

*--ht-alt-liftover*の代わりに*--ht-alt-aware-validate false*オプションを使用できます。ただし、hg19.faリファレンスに代替コンティグが存在するため、dragenマッピングの品質が大きな影響を受けます。

エピゲノムマッピング/アライメント：ディレクショナルプロトコール、シングルエンドのFASTQインプット、BAMアウトプット

ディレクショナル（Lister）プロトコールは、バイサルファイトシーケンシングストランド候補4つのうち2つからリードを作成します。したがって、*--methylation-protocol=directional*オプションを使用した場合、DRAGENは、2つのストランド候補に対応する制約に対して、各リードまたはリードペアを2回アライメントします。以下のDRAGENコマンドは、2つの異なるBAMファイルを作成します：

```
mkdir -p /staging/epigenome/directional
dragen -f -r /staging/human/reference/hg19_epigenome \
-1 /staging/epigenome/reads/sample_1_R1.fastq.gz \
-2 /staging/epigenome/reads/sample_10_R2.fastq.gz \
--RGID Illumina_RGID \
--RGSM sample_10 \
--RGPL illumina \
--output-directory /staging/epigenome/directional \
--output-file-prefix sample_10 \
--methylation-protocol=directional \
--enable-sort false
```

エピゲノムマッピング/アライメント：ノンディレクショナルプロトコール、ペアエンドのFASTQインプット、BAMアウトプット

ノンディレクショナルプロトコールは、4つのバイサルファイトシーケンシングストランド候補すべてからリードを作成します。このため、*--methylation-protocol=non-directional*引数を使用した場合、DRAGENは各リードを4回アライメントし、4つの異なるBAMファイルを作成します。

```
mkdir -p /staging/epigenome/non-directional
dragen -f -r /staging/human/reference/hg19_epigenome \
-1 /staging/epigenome/reads/sample_10_R1.fastq.gz \
-2 /staging/epigenome/reads/sample_10_R2.fastq.gz \
--RGID Illumina_RGID \
--RGSM sample_10 \
--RGPL illumina \
--output-directory /staging/epigenome/non-directional \
--output-file-prefix sample_10 \
--methylation-protocol non-directional \
--enable-sort false
```

バリアントコーリングのみの例

バリアントコーリングのみの例は、既存のアライメント済みBAMまたはCRAMファイルを直接、DRAGENバリアントコーラーに渡す方法を示します。初期設定では、BAM/CRAMファイルは、バリアントコーリングの前、ソーティング段階で渡されます。

DRAGENバリアントコーラーを実行する前に、BAMファイルを重複マーキングするには、別のツールを使用する必要があります。DRAGEN Duplicate Markerは、BAMファイル内に存在しないマッパー/アライナーの提供する情報に依存します。DRAGEN Duplicate Markerを活用するには、DRAGENをエンドツーエンドモードで使用します。

これらのコマンド例でインプットとして使用されるBAM/CRAMファイルは、サンプルデータセットには含まれていません。これらは、28 ページの 「アライメントのみの例」 にあるコマンド例を使って生成することができます。

未ソートBAMインプット、VCFアウトプット（初期設定）

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-b /staging/examples/unsorted_SRA056922_30x_e10_50M.bam \
--enable-variant-caller true \
--output-directory /staging/examples/ \
--output-file-prefix unsorted_output_SRA056922_30x_e10_50M \
--enable-map-align false
```

- または、`/opt/edico/examples/unsorted_bam_in_vcf_out.sh`を実行します。

ソート済みBAMインプット、VCFアウトプット

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-b /staging/examples/SRA056922_30x_e10_50M.bam \
--enable-variant-caller true \
--output-directory /staging/examples/ \
--output-file-prefix sorted_output_SRA056922_30x_e10_50M \
--enable-map-align false
--enable-sort false
```

- または、`/opt/edico/examples/sorted_bam_in_vcf_out.sh`を実行します。

ソート済みCRAMインプット、VCFアウトプット

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
--enable-variant-caller true \
--output-directory /staging/examples/ \
--output-file-prefix sorted_output_SRA056922_30x_e10_50M \
```

```
--enable-sort false \
--enable-map-align false \
--cram-input /staging/examples/SRA056922_30x_e10_50M.cram
```

- または、/opt/edico/examples/sorted_cram_in_vcf_out.shを実行します。

Somaticスモールバリアントコーラーの例

Tumor-Normal BAMインプットオプションを使用しているときに、BAMリードグループがRGIDを共有している場合、DRAGENは、リードが所属するリードグループを判断できません。理想的には、リードグループごとに異なるRGIDを持つべきですが、*--prepend-filename- to-rgid*をtrueに設定して、この問題を回避することができます。

ペアエンドFASTQインプット

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
--tumor-fastq1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_
1.fastq.gz \
--tumor-fastq2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_
2.fastq.gz \
--enable-variant-caller true \
--RGID-tumor DRAGEN_RGID \
--RGSM-tumor DRAGEN_RGSM \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M
```

ソート済みBAMインプット

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
--tumor-bam-input /staging/examples/SRA056922_30x_e10_50M.bam \
--enable-variant-caller true \
--output-directory /staging/examples/ \
--output-file-prefix sorted_output_SRA056922_30x_e10_50M \
--enable-map-align false \
--prepend-filename-to-rgid true
```

gVCFとジェノタイピングの例

ペアエンドFASTQインプット、gVCFアウトプット

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--enable-variant-caller true \
--vc-emit-ref-confidence GVCF \
--RGID Illumina_RGID \
--RGSM SRA056922_30x_e10_50M \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M
```

- または、/opt/edico/examples/paired_fastq_in_gVCF_out.shを実行します。

gVCFインプットによるジョイントコール

- 以下のとおり入力します：

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
--enable-joint-genotyping true \
--output-directory /staging/examples/ \
--output-file-prefix Joint_SRA056922_30x_e10_50M \
--variant /staging/examples/SRA056922_30x_e10_50M.gvcf.gz
```

- または、/opt/edico/examples/single_gVCF_in_jointVCF_out.shを実行します。

Pedigreeベースのジョイントジェノタイピング：3つのgVCFファイルと1つのpedigreeファイルインプット、ジョイントジェノタイプ済みVCFアウトプット

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
--enable-joint-genotyping true \
--output-directory /staging/examples/ \
--output-file-prefix Joint_SRA056922_30x_e10_50M \
--variant /staging/examples/mother.gvcf.gz \
--variant /staging/examples/father.gvcf.gz \
--variant /staging/examples/child.gvcf.gz \
--pedigree-file <PEGIGREE_FILE>
```

1ステップ集団ベースのジョイントジェノタイピング：gVCFインプット、ジョイントジェノタイプ済みマルチサンプルアウトプット

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
--enable-joint-genotyping true \
```

```
--output-directory /staging/examples/ \
--output-file-prefix Joint_SRA056922_30x_e10_50M \
--variant /staging/examples/SRA056922_30x_e10_50M.gvcf.gz
```

2ステップ集団ベースのジョイントジェノタイピング：gVCFリストインプット、ジョイントジェノタイプ済みマルチサンプルVCFアウトプット

第1ステップでは、アウトプットとして、マルチサンプルVCFが生成されます。このステップでは、gVCFリストをインプットとする以下のコマンドラインオプションが使用されます。

```
dragen -f \
--enable-gvcf-genotyper true \
--enable-map-align false \
--variant-list ${GVCF_LIST} \
--ht-reference ${FASTA_REF} \
--intermediate-results-dir ${TEMP_DIR} \
--output-directory ${OUTPUT_DIR} \
--output-file-prefix ${COHORT_NAME}
```

第2ステップでは、ジョイントジェノタイプ済みのマルチサンプルVCFが生成されます。このステップでは、第1ステップで作成されたマルチサンプルVCFをインプットとする以下のコマンドラインオプションが使用されます。マルチサンプルVCFを指定するには、*-- variant*を使用します。

```
dragen -f \
--enable-joint-genotyping true \
--variant ${MULTISAMPLE_VCF} \
--ref-dir ${FASTA_REF} \
--output-directory ${OUTPUT_DIR}\
--output-file-prefix ${COHORT_NAME}.joint_genotyped
```

表 2　ジョイントコーリングモードと関連する入力ファイル、およびコマンドラインオプション

生成するVCF	集団ジョイントコールマルチサンプルgVCF	ファミリージョイントコールマルチサンプルgVCF	集団ジョイントコールマルチサンプルVCF	ファミリージョイントコールマルチサンプルVCF
入力ファイル	マルチサンプル統合gVCFファイル	マルチサンプル統合gVCFファイル	マルチサンプル統合gVCFファイルまたはX個の独立したgVCFファイル	マルチサンプル統合gVCFファイルまたはX個の独立したgVCFファイル
pedigreeファイルの使用	使用しない	使用する	使用しない	使用する
コマンドラインオプション	--enable-joint-genotyping true--enable-multi-sample-gvcf=TRUE	--enable-joint-genotyping true --enable-multi-sample-gvcf=TRUE --pedigree-file file.ped	--enable-joint-genotyping true	--enable-joint-genotyping true --pedigree-file file.ped

カバレッジメトリクスレポートの例

DRAGENの初期設定では、全ゲノムに対するリードカバレッジがレポートされます。また、もし利用可能であれば、ターゲットBEDもレポートの対象となります。カバレッジのレポート領域は3つまで指定できます。以下の例では、DRAGENにより、`cov_report`と、追加のカバレッジ領域1用の`full_res`の2つのカバレッジレポートが作成されます。

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--RGID Illumina_RGID \
--RGSM SRA056922_30x_shuffle16k \
--output-directory /staging/examples/ \
--output-file-prefix SRA056922_30x_e10_50M \
--qc-coverage-region-1 /staging/examples/reads/vc_smoke.callable.bed \
--qc-coverage-reports-1 cov_report full_res
```

`full_res`レポートは、Bedtoolsカバレッジオプションに対応し、塩基1つあたりの解像度リード深度を含みます。`cov_report`は、深度の平均値、中央値、最大値、最小値など、領域1つあたりのリード深度サマリーを含みます。

CNVの例

これらの例は、すでにマッピングおよびアライメントの済んでいるBAMファイルを、DRAGEN CNVを使って処理する方法を示しています。DRAGEN CNVパイプラインはセルフノーマライゼーションと正常サンプルのパネルの2つのモードをサポートしています。

セルフノーマライゼーションには、*enable-cnv=true*オプションを使って、DRAGENハッシュテーブルを生成する必要があります。CNVを頻繁に実行する場合は、必ず、CNV対応ハッシュテーブルを生成することを推奨します。

*enable-map-align*オプションは、初期設定では、構成ファイルでtrueに設定されています。インプットBAMのマッピングとアライメントが必要ない場合は、falseに設定します。

*--intermediate-results-dir*オプションにはローカルディレクトリ（例：`/staging/intermediate`または`/local` ssd）を設定する必要があります。そうしないと、CNVステップでの処理時間が長くなる可能性があります。

セルフノーマライゼーションを使った実行

単一サンプルのWGS処理には、セルフノーマライゼーションが好んで行われます。BAMは、コマンドライン1つで、CNVパイプライン全体を経ます。

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-b /staging/examples/SRA056922_30x_e10_50M.bam \
--intermediate-results-dir /staging/intermediate \
--output-directory /staging/examples/ \
--output-file-prefix dragen_cnv1 \
--enable-map-align false \
--enable-cnv true \
--cnv-enable-self-normalization true \
```

正常サンプルのパネルを使った実行

正常サンプルのパネルを使ったアプローチでは、使用するサンプルそれぞれに対してtarget.countsファイルをあらかじめ生成しておく必要があり、最後に1つコマンドを実行して、ノーマライゼーションとコピー数バリアントコールを行います。

BAMインプットを使ってターゲットカウントを計算するために、このサンプルコマンドは、BAMファイルのアライメントから、リードカウントを含むシグナルを抽出し、ノーマライゼーションステップで使用される*.target.countsファイルを生成します。ターゲットカウントは、解析中の腫瘍サンプルと正常サンプルを含め、インプットBAMファイルごとに計算する必要があります。

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-b /staging/examples/SRA056922_30x_e10_50M.bam \
--intermediate-results-dir /staging/intermediate \
--output-directory /staging/examples/ \
--output-file-prefix dragen_cnv1 \
--enable-map-align false \
--enable-cnv true
```

以下のコマンドは、ノーマライゼーションを実行し、CNVコールを生成します。正常サンプルは、正常サンプルの*.target.countsファイルへのパスを提供するテキストファイル（この例では`normal.txt`）に列挙されます。ケースサンプルの*.target.countsファイルの指定には、*-- cnv-input*オプションを使用します。この例では、インプットのgcbias修正は無効化されています。

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
--intermediate-results-dir /staging/intermediate \
--output-directory /staging/examples/ \
--output-file-prefix dragen_cnv2 \
--enable-cnv true \
--cnv-input /staging/examples/dragen_cnv1.target.counts \
--cnv-normals-list normal.txt \
--cnv-enable-gcbias-correction false
```

FASTQ処理

この例は、FASTQサンプルから直接、セルフノーマライゼーションモードでDRAGEN CNVコーラーを実行します。まず、FASTQのマッピングとアライメントを行い、続けて、直接、CNVコールします。このステップは、バリアントコーリングと組み合わせることができます。

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_1.fastq.gz \
-2 /staging/examples/reads/SRA056922_30x_shuffle16k_e10_50M_2.fastq.gz \
--RGID Illumina_ID \
--RGSM SRA056922_30x_shuffle16k \
--intermediate-results-dir /staging/intermediate \
--output-directory /staging/examples/ \
--output-file-prefix dragen_cnv \
--enable-map-align true \
--enable-cnv true \
--cnv-enable-self-normalization true
```

De Novo CNVコールの実行

De Novoコールには、過去に単一サンプル解析で生成したノーマライズ済みのシグナルファイル（*.tn.tsv）が必要です。pedigreeファイルが提供されている場合、de novo状態とde novoクオリティスコアが、発端者サンプルのレコードにアノテーションされます。

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
--cnv-input father.tn.tsv \
--cnv-input mother.tn.tsv \
--cnv-input child.tn.tsv \
--intermediate-results-dir /staging/intermediate \
--output-directory /staging/examples/ \
--output-file-prefix trio_cnv \
--pedigree-file trio.ped \
--enable-cnv true
```

体細胞CNVコールの実行

体細胞CNVコールには、腫瘍サンプルおよびそれにマッチする正常サンプルが必要です。まず、マッチする正常サンプルを生殖細胞系列スモールバリアントコーラーに通して、*.hard-filtered.vcf.gzを作成する必要があります。もし、わかっていれば、サンプルの性別を指定することを推奨します。

```
dragen -f \
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
--tumor-bam-input tumor.bam \
--intermediate-results-dir /staging/intermediate \
```

```
--output-directory /staging/examples/ \
--output-file-prefix somatic_cnv \
--enable-map-align false \
--enable-cnv true \
--cnv-normal-b-allele-vcf normal.hard-filtered.vcf.gz \
--sample-sex female
```

構造多型コールの例

構造多型コールは次のモードで実行できます：

- スタンドアロン：マッピングされたBAM/CRAM入力ファイルから実行されます。*--enable-map-align=false*オプションと*--enable-sv=true*オプションが必要です。
- 統合：DRAGENマッパー/アライナーの出力で自動的に実行されます。*--enable-map-align=true*、*--enable-sv=true*、*--enable-map-align-output=true*オプションが必要です。

構造多型コールは、他のコーラーとともに有効化できます。

統合実行の例

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

スタンドアロンのジョイント二倍体コールの例

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

スタンドアロンのDe Novoクオリティスコアリングの例

```
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>
```

最小設定でのBCLからFASTQへの変換

この例では、DRAGENを使用して、Illumina BCL形式のファイルを処理する方法を示します。

この例で使用されるBCLディレクトリは、サンプルデータパッケージには含まれていません。`/mnt/san/131022_hsxten008_0123_FC543`をお使いのBCLディレクトリで置き換えてください。

- 以下のとおり入力します：

  ```
  dragen --bcl-conversion-only=true \
  --bcl-input-directory /mnt/san/131022_hsxten008_0123_FC543 \
  -- output-directory /staging/examples/
  ```

- または、`/opt/edico/examples/bcl_in_fastq_out.sh`を実行します。

S3およびHTTPストリームインプットの例

DRAGENは、S3バケットから直接、または、インプットストリームとして知られているHTTP署名付きURLから直接、入力ファイルを処理できます。処理の前に、入力ファイルをローカルディスクにダウンロードしておく必要はありません。代わりに、それらのファイルが、ネットワーク経由で直接、DRAGENプロセッサーへストリームされます。

ストリームは、圧縮されたFASTQ（*.fastq.gz）ファイルでサポートされています。DRAGENの今後のバージョンでは、BAM（*.bam）ファイルからのストリームもサポートされる予定です。インプットストリームは、シングルエンドFASTQ、ペアエンドFASTQ、およびFASTQリストを使用するすべての構成で使用できます。以下の例で、インプットストリームの使い方をいくつか紹介します。

S3を使用したFASTQインプットのストリーム

```
dragen -f
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 s3://s3-bucket-name/path/to/object_1.fastq.gz \
-2 s3://s3-bucket-name/path/to/object_2.fastq.gz \
--RGID object_ID \
--RGSM sample_name \
--output-directory /staging/examples/ \
--output-file-prefix streaming
```

HTTPを使用したFASTQインプットのストリーム

```
dragen -f
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 https://bucket-name.amazonaws.com/path/to/object_1.fastq.gz?querystring
\
-2 https://bucket-name.amazonaws.com/path/to/object_2.fastq.gz$querystring
\
--RGID object_ID \
-RGSM sample_name \
--output-directory /staging/examples/ \
--output-file-prefix streaming
```

リモートファイルにアクセスするには許可を得る必要があります。そのファイルへのアクセス権を持っていれば、DRAGENでリモートファイルをストリームできます。S3オブジェクトでは、AWS認証と認証情報が必要です。AWS認証は、例えば、IAMポリシーなどを使って、実行するインスタンスにあらかじめ設定しておく必要があります。HTTP URLには、ほとんどの場合、クエリ文字列が添付されていて、この文字列に、認証情報、または許可を与えるために必要なトークンが含まれています。セキュリティの手段がURLのほかの部分に存在することもあります。例えば：

```
https://stagingdl.dnanex.us/security/string/sample_1.fastq.gz
```

マルチコーラーワークフロー

DRAGENでは、単一のワークフローで複数のツールを実行できます。

*enable-component*フラグは、コンポーネントの有効化、無効化を制御します。DRAGENは、有効化されたコンポーネントを使ってワークフローを構築し、コンポーネントの不整合を自動的に解決します。可能であれば、DRAGENはコンポーネントを並行実行します。

コンポーネントはそれぞれ、インプット設定や内部アルゴリズムパラメーター、出力ファイルやフィルター条件などの設定するための複数のオプションを持っています。詳細については、各コンポーネントのセクションを参照してください。

*output-directory*や*sample-sex*など、一部のオプションは、複数のコーラーで共有されます。

各バリアントコーラーは、VCFとメトリクス出力ファイルのセットを独自に作成します。

コンポーネントコマンドの例

```
enable-map-align
enable-sort
enable-duplicate-marking
enable-variant-caller
enable-cnv
enable-sv
```

インプット形式

DRAGENが受け入れる一般的かつ標準的なNGSインプット形式は以下のとおりです：

- FASTQ（`fastq-file1`および`fastq-file2`）
- FASTQ List（`fastq-list`）
- BAM（`bam-input`）
- CRAM（`cram-input`）

体細胞ワークフローは、腫瘍に相当する入力ファイル（`tumor-bam-input`など）を使用できます。

アライメントされていないリードから実行した場合、リードはまず、マッピング/アライメントコンポーネントを通って、アライメントを作成します。このアライメントはそのまま下流のバリアントコーラーまで進みます。あらかじめアライメントされたリードから実行した場合、DRAGENでは、マッピング/アライメントコンポーネントを使って再アライメントする、またはソース入力からの既存のアライメントを使用することができます。

マルチコーラーコマンドラインの例

ここでは、シングルコーラーを使用した場合のコマンドラインオプションを組み合わせで、マルチコーラーワークフローを作成する際のベストプラクティスの一例を示します。この例は、以下のステップから構成されます：

- INPUTオプションを設定。
- OUTPUTオプションを設定。
- 再アライメントの有無に応じてMAP/ALIGNを設定。
- 用途に基づいてバリアントコーラーを設定。
- コンポーネントごとに必要なオプションを構築し、最後のコマンドラインで再利用を可能に。

```
INPUT_OPTIONS="
--ref-dir $DRAGEN_HASH_TABLE \
--fastq-file1 $FASTQ1 \
--fastq-file2 $FASTQ2 \
--RGSM $RGSM \
--RGID $RGID \
"
OUTPUT_OPTIONS="
--output-directory $OUTPUT \
--output-file-prefix $PREFIX \
"
MA_OPTIONS="
--enable-map-align true \
... <any other optional settings> \
"
CNV_OPTIONS="
```

```
    --enable-cnv true \
    ... <any other optional settings> \
    "
    SNV_OPTIONS="
    --enable-variant-caller true \
    ... <any other optional settings> \
    "
    SV_OPTIONS="
    --enable-sv true \
    ... <any other optional settings> \
    "
    CMD="
    dragen \
    $INPUT_OPTIONS \
    $OUTPUT_OPTIONS \
    $MA_OPTIONS \
    $CNV_OPTIONS \
    $SNV_OPTIONS \
    $SV_OPTIONS \
    "
```

生殖細胞系列

以下の表に、サポートされているインプット形式とバリアントコーラーの一部をまとめます。この表は抜粋で、サポートされていても記載されていない機能やコーラーもあります。

生殖細胞系列	マッピング/アライメントを使ったFASTQ	BAM/CRAM	マッピング/アライメントを使ったBAM/CRAM
CNV + SNV	サポート	サポート	サポート
CNV + SV	サポート	サポート	サポート
SNV + SV	サポート	サポート	サポート
CNV + SNV + SV	サポート	サポート	サポート

体細胞

体細胞ワークフローは、腫瘍インプットと正常インプットの両方を指定します。さらに体細胞CNVコーラー用に2つの入力ファイル（腫瘍とマッチする正常）が必要であるだけでなく、マッチする正常SNV VCFがが必要になる可能性があるということは、細心の注意を払う必要があることを意味します。このため、推奨されるTumor-Normalワークフローでは、まず、マッチする正常インプットを生殖細胞系列ワークフローで解析します。

1. マッチする正常インプットを、生殖細胞系列ワークフロー（CNV + SNV + SV + ...）で解析します。このワークフローにより、マッチする正常SNV VCFが生成されます。
2. 腫瘍インプットとマッチする正常インプットを、体細胞ワークフロー（CNV + SNV + SV + ...）で解析します。

```
INPUT_OPTIONS="
--ref-dir $DRAGEN_HASH_TABLE \
--tumor-bam-input $TUMOR_BAM \
--bam-input $NORMAL_BAM \
"
OUTPUT_OPTIONS="
--output-directory $OUTPUT \
--output-file-prefix $PREFIX \
"
MA_OPTIONS="
--enable-map-align false \
... <any other optional settings> \
"
CNV_OPTIONS="
--enable-cnv true \
--cnv-normal-b-allele-vcf $SNV_VCF \
... <any other optional settings> \
"
SNV_OPTIONS="
--enable-variant-caller true \
... <any other optional settings> \
"
SV_OPTIONS="
--enable-sv true \
... <any other optional settings> \
"
CMD="
dragen \
$INPUT_OPTIONS \
$OUTPUT_OPTIONS \
$MA_OPTIONS \
$CNV_OPTIONS \
$SNV_OPTIONS \
$SV_OPTIONS \
"
```

以下の表は、Tumor-Normalモードでサポートされているさまざまな組み合わせをまとめたものです。

Tumor-Normal	マッピング/アライメントを使ったFASTQ	BAM/CRAM	マッピング/アライメントを使ったBAM/CRAM
CNV + SNV	サポート	サポート	未サポート
CNV + SV	サポート	サポート	未サポート
SNV + SV	サポート	サポート	未サポート
CNV + SNV + SV	サポート	サポート	未サポート

Tumor-onlyモードで実行するには、INPUTオプションからマッチする正常インプットを削除し、Tumor-onlyモードで実行するように、個々のコーラーを設定します。以下の表は、Tumor-onlyモードでサポートされているさまざまな組み合わせをまとめたものです。

Tumor-only	マッピング/アライメントを使ったFASTQ	BAM/CRAM	マッピング/アライメントを使ったBAM/CRAM
CNV + SNV	サポート	サポート	サポート
CNV + SV	サポート	サポート	サポート
SNV + SV	サポート	サポート	サポート
CNV + SNV + SV	サポート	サポート	サポート

WES解析は、そのモードがシングルコーラーモードでサポートされていて、インプット設定に不一致がない場合にサポートされます。

DRAGENホストソフトウェア

DRAGENホストソフトウェアプログラム*dragen*を使って、リファレンスゲノムを構築し、ロードしてから、シーケンスデータを解析することができます。このとき、データの展開、マッピング、アライメント、ソート、除去オプションも使用可能な重複マーキング、バリアントコーリングが行われます。

このソフトウェアを起動するには、*dragen*コマンドを使用します。これ以降のセクションでは、コマンドラインオプションについて説明します。

コマンドラインオプションは構成ファイルでも設定できます。構成ファイルの詳細については、64 ページの 「構成ファイル」 を参照してください。同じオプションが、構成ファイルとコマンドラインの両方で指定されている場合、コマンドラインオプションが優先されます。

コマンドラインオプション

コマンドラインオプションの一覧については、48 ページの 「コマンドラインオプション」 を参照してください。

DRAGENコマンドラインオプション

dragenを使って、以下のコマンドラインオプションを実行します：

- リファレンス/ハッシュテーブルの構築

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

- マッピング/アライメントおよびバリアントコーラーの実行（**.fastqから*.vcf*へ）

```
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
```

- マッピング/アライメントの実行（**.fastqから*.bam*へ）

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

- バリアントコーラーの実行（**.bamから*.vcf*）

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

- BCLコンバーターの実行（*BCLから*.fastq*）

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

- RNAマッピング/アライメントの実行（**.fastqから*.bam*）

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

リファレンスゲノムオプション

リードのアライメントにDRAGENシステムを使用するには、まず、リファレンスゲノムと関連するハッシュテーブルをPCIeカードにロードする必要があります。リファレンスゲノムのFASTAファイルを前処理して、DRAGENのネイティブバイナリーリファレンスとハッシュテーブルの形式に変換する方法について、詳しくは、10 ページの 「リファレンスゲノムの準備」 を参照してください。また、*-r [or --ref-dir]*オプションを使って、前処理されたバイナリーリファレンスとハッシュテーブルを含むディレクトリを指定する必要もあります。この引数は常に必須です。

リードの処理とは別に、以下のコマンドを使用して、リファレンスゲノムとハッシュテーブルをDRAGENカードメモリーにロードします。

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

リファレンスゲノムがすでにロードされている場合でも、*-l (--force-load-reference)*オプションを使用して、強制的にロードします。

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

リファレンスゲノムのロードに必要な時間は、リファレンスのサイズによって異なりますが、一般に推奨される設定では、約30～60秒です。

出力オプション

出力については、以下のコマンドラインオプションが必須です：

- *--output-directory <out_dir>*：生成されたファイルの出力先ディレクトリを指定します。
- *--output-file-prefix <out_prefix>*：出力ファイルの接頭辞を指定します。DRAGENは生成されたファイルそれぞれに対して、この接頭辞に適切なファイル拡張子を追加します。
- *-r [--ref-dir]*：リファレンスハッシュテーブルを指定します。

以下の例には、これらの必須オプションは含まれていません。

マッピングとアライメントの場合、初期設定では、出力はソートされ、BAM形式で圧縮されてから、ディスクに保存されます。マッピング/アライメント段階で出力形式を指定するには、*--output-format <SAM|BAM|CRAM>*オプションを使用します。出力ファイルが存在する場合、DRAGENは警告メッセージを表示して、終了します。既存の出力ファイルを強制的に上書きする場合は、 *-f [--force]*オプションを使用します。

例えば、以下のコマンドは、圧縮されたBAMファイルに出力し、強制的に上書きします：

```
dragen ... -f
dragen ... -f --output-format bam
```

BAI形式のBAMインデックスファイル（*.bai）を生成するには、*--enable-bam-indexing*をtrueに設定します。

以下の例は、SAMファイルに出力し、強制的に上書きします：

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

以下の例は、CRAMファイルに出力し、強制的に上書きします：

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

DRAGENは、BAM標準で説明されている通りのミスマッチ差分（MD）タグを生成できます。この機能は、初期設定ではオフになっています。これは、この文字列の生成には多少のパフォーマンスコストがかかるからです。MDタグを生成するには、*--generate-md-tags*をtrueに設定します。

ZS:Zアライメントステータスタグを生成するには、*--generate-zs-tags*をtrueに設定します。これらのタグは、一次アライメントで、リードが二次出力となれる準最適アライメントを持つときにのみ（*--Aligner.sec-aligns*が0に設定されていて、何も出力されなかった場合でも）、生成されます。

有効なタグの値は以下のとおりです：

- ZS:Z:R：類似したスコアを持つ複数のアライメントが見つかった。
- ZS:Z:NM：アライメントが見つからなかった。
- ZS:Z:QL：アライメントは見つかったが、クオリティ閾値を下回った。

SA:Zタグを生成するには、*--generate-sa-tags*をtrue（初期設定）に設定します。これらのタグは、補足的アライメントグループのアライメント情報（位置、CIGAR、方向）を提供しますが、これは構造多型コールで有用です。

BQSRタグの保持と除去

Picard Base Quality Score Recalibration（BQSR）ツールは、BIタグとBDタグを含む出力BAMファイルを作成します。BQSRは、リードのシーケンスに対し正確にタグを計算します。BIタグとBDタグを持つBAMファイルがマッパー/アライナーへのインプットとして使用されているときに、ハードクリッピングが有効化されている場合、BIまたはBDタグ、もしくはその両方が無効になります。

BAMファイルを入力として使用する場合は、これらのタグを取り除くことを推奨します。BIタグとBDタグを取り除くには、*--preserve-bqsr-tags*オプションをfalseに設定します。これらのタグをそのままにしておいた場合、DRAGENから、ハードクリップの無効化を求める警告が発せられます。

リードグループオプション

DRAGENは、特定のFASTQにあるリードはすべて、同じリードグループに属することを前提にしています。DRAGENは、出力BAMファイルのヘッダーに、以下の標準BAM属性を指定する能力を持つ@RGリードグループディスクリプターを1つ作成します：

属性	引数	説明
ID	--RGID	リードグループ識別子。リードグループパラメーターのいずれかを指定する場合、RGIDは必須です。この値は、各出力BAMレコードに書き込まれます。
LB	--RGLB	ライブラリー。

属性	引数	説明
PL	--RGPL	リードの作成に使用されたプラットフォーム/テクノロジー。BAM標準で使用が許されている値は、CAPILLARY、LS454、ILLUMINA、SOLID、HELICOS、IONTORRENT、PACBIOです。
PU	--RGPU	プラットフォームのユニット。例:`flowcell-barcode.lane`。
SM	--RGSM	サンプル。
CN	--RGCN	リードを作成したシーケンシングセンターの名前。
DS	--RGDS	説明。
DT	--RGDT	ランの作成日。
PI	--RGPI	予測される平均値のインサートサイズ。

これらの引数のいずれかが存在する場合、DRAGENは出力レコードすべてにRGタグを追加して、それらが同じリードグループのメンバーであることを示します。以下に、リードグループパラメーターを含むコマンドラインの例を示します：

```
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
```

複数のリードグループの入力に、*--fastq-list*オプションを使用した場合、`fastq_list.csv`ファイルに列を追加することにより、各リードグループにBAMタグ（など）が指定されます。列のヘッダーはそれぞれ、「RG」で始まる4つの大文字で構成されます。それぞれの列について、その列の各リードグループの値が、同じ名前のタグに入れられて、出力BAMファイルに追加されます。

ライセンスオプション

ランの最後に、ライセンスのステータスメッセージが表示されないようにするには、*--lic-no-print*オプションを使用します。ライセンスステータスメッセージの例を以下に示します：

```
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
```

動作モード

DRAGENの主たる動作モードは以下の2つです：

- マッパー/アライナー
- バリアントコーラー

DRAGENでは、それぞれのモードを個別に実行することもできますし、エンドツーエンドのソリューションとして実行することもできます。また、DRAGENパイプラインに沿って、展開、ソーティング、重複マーキング、圧縮の有効/無効を切り替えることも可能です。

フルパイプラインモード

フルパイプラインモードを実行するには、*--enable-variant-caller*をtrueに設定し、インプットを、*.fastq、*.bam、または*.cram形式のマッピングされていないリードとして提供します。

DRAGENは、展開、マッピング、アライメント、ソーティング、および必要に応じて重複マーキングを行い、直接、バリアントコーラーにかけて、VCFファイルを作成します。このモードでは、DRAGENは、パイプライン全体を通じて並行ステージを使用し、ランタイムを大幅に削減します。

マッピング/アライメントモード

マッピング/アライメントモードは初期設定で有効になっています。インプットは、*.fastq、*.bam、または*.cram形式のマッピングされていないリードです。DRAGENは、アライメントされたソート済みのBAMまたはCRAMファイルを作成します。同時に重複リードをマーキングするには、*--enable-duplicate-marking*を*true*に設定します。

バリアントコーラーモード

バリアントコーラーモードを実行するには、*--enable-variant-caller*オプションをtrueに設定します。

インプットは、マッピングされ、アライメントされたBAMファイルです。DRAGENは、VCFファイルを作成します。BAMファイルがソート済みの場合は、*--enable-sort*を*false*に設定して、ソーティングをスキップできます。BAMファイルが重複マーキングされていない場合、バリアントコーリングの前に、DRAGENパイプラインでマーキングすることはできません。重複マーキング機能を活用するには、エンドツーエンドで操作モードを使用します。

RNA-Seqデータ

RNA-Seqベースのデータの処理を可能にするには、*--enable-rna*を*true*に設定します。

DRAGENは、マッパー/アライナーステージで、RNAスプライスアライナーを使用します。DRAGENは、操作に必要なモードを動的に切り替えます。

バイサルファイトMethylSeqデータ

バイサルファイトMethylSeqデータの処理を可能にするには、*--enable-methylation-calling*オプションをtrueに設定します。DRAGENは、Lister（ディレクショナル）プロトコールとCokus（ノンディレクショナル）プロトコールのデータ処理を自動化して、bismark互換タグを持つBAMファイルを1つ生成します。

また、代わりに、C->TおよびG->A変換されたリードおよびリファレンスの組み合わせそれぞれに対して個別にBAMファイルを作成するモードでDRAGENを実行することもできます。この処理モードを有効化するには、*--ht-methylated*を有効にして、リファレンスハッシュテーブルを構築し、適切な*--methylation-protocol*設定を使ってDRAGENを実行する必要があります。

入力オプション

DRAGENは、FASTQ形式、またはBAM/CRAM形式のリードを処理できます。FASTQ入力ファイルには、以下の圧縮オプションを指定できます。

- 圧縮なし
- gzipまたはbgzip圧縮
- ORA圧縮。ORA圧縮を使用するには、ORAリファレンスとリファレンスディレクトリを提供する必要があります。346 ページの 「DRAGEN ORA圧縮と展開」 を参照してください。

FASTQ入力ファイルがgzipされている場合、ハードウェアアクセラレーションによる展開を使用して、ファイルが自動的に展開され、その後、リードがマッパーにストリームされます。ファイルの拡張子が*.oraである場合、DRAGEN はORA展開を使用して、ファイルが自動的に展開され、その後、リードがマッパーにストリームされます。これらのFASTQコマンドラインオプションは、すべての圧縮形式で利用できます。

入力ファイルの種類

使用する入力ファイルに応じて、以下のコマンドラインオプションを使用します。

FASTQ入力ファイル

FASTQ入力ファイルは、シングルエンド、またはペアエンドです。FASTQファイルをインプットするには、以下の例を使用します。

- **FASTQファイル1つのシングルエンド**（*-1オプション*）

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

- **マッチするFASTQファイル2つのペアエンド**（*-1と-2オプション*）

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

- **1つのマージされたFASTQファイルのペアエンド**（*--interleaved (-i)オプション*）

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

bcl2fastqまたはDRAGEN BCLコマンドを使用する場合、以下の命名規則を使用します：

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

bcl2fastqとDRAGENの以前のバージョンでは、FASTQサンプルを複数のファイルセグメントに分割して、ファイルサイズを制限したり、生成時間を短縮したりすることができました。

例えば：

```
RDRS182520_S1_L001_R1_001.fastq.gz
RDRS182520_S1_L001_R1_002.fastq.gz
...
RDRS182520_S1_L001_R1_008.fastq.gz
```

これらのファイルは、結合しなくても、DRAGENでまとめて処理できます。サンプルをマッピング/アライメントするには、先頭ファイルを指定します（*-1 <FileName>_001.fastq*）。DRAGENは、*-1*および*-2*オプションを使って指定されたペアエンドインプットおよびfastq.gz圧縮ファイルの、サンプル内すべてのセグメントファイルを連続的に読み込みます。この動作をオフにするには、コマンドラインで、*--enable-auto-multifile*をfalseに設定します。

また、DRAGENでは、必要に応じて、ファイル名に指定されたサンプル名に従って複数のファイルを読み込むこともできます。この方法を使用すると、複数のBCLレーンまたはフローセルにわかれているサンプルを組み合わせることができます。この機能を有効にするには、*--combine-samples-by-name*オプションをtrueに設定します。

コマンドラインで指定されたFASTQファイルが、前述のCasava 1.8命名規則を使用し、同じディレクトリにあるその他のファイルが同じサンプル名を共有している場合、これらのファイルとすべてのセグメントが自動的に処理されます。ただし、サンプル名、リード番号、ファイル拡張子は一致しなければなりません。インデックスバーコードとレーン番号は違っていてもかまいません。

システムパフォーマンスへの影響を回避するには、入力ファイルを、高速ファイルシステムに入れる必要があります。

複数のFASTQ入力ファイル

複数のFASTQ入力ファイルを指定するには、*--combine- samples-by-name*オプションの代わりに、*--fastq-list <csv file name>*オプションを使って、FASTQファイルのリストを含むCSVファイルの名前を指定することを推奨します。例えば：

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

CSVファイルを使用すると、FASTQ入力ファイルの名前を指定し、複数のサブディレクトリから入力して、各リードグループに対して明確に指定されたBAMタグを追加することができます。DRAGENは、FASTQへのBCL変換中、正しい形式でCSVファイルを自動生成します。このCSVファイルには`fastq_list.csv`という名前が付けられ、その中には、ラン中に作成されたFASTQファイル1つ、またはペアエンドファイルのペア1つにつき1つのエントリーが含まれます。

FASTQ CSVファイル形式

CSVファイルの先頭行は、各列のタイトルで、その下にデータ行が1行以上続きます。CSVファイル1行に含まれるカンマで区切られた値の数は、すべての行で同じです。また、スペースなどの関係のない文字が含まれていてはいけません。

列のタイトルでは大文字と小文字が区別されます。以下の列タイトルが必須です；

- RGID：リードグループ
- RGSM：サンプルID
- RGLB：ライブラリ
- Lane：フローセルレーン
- Read1File：有効なFASTQ入力ファイルへのフルパス
- Read2File：有効なFASTQ入力ファイルへのフルパス。ペアエンドインプットでは必須です。ペアエンドインプットを使用しない場合は、空欄のままにしてください。

1つのFASTQファイルがCSVリストで参照されるのは1回だけです。Read2File列の値はすべて、空ではなく、有効なファイルを参照しているか、またはすべて空でなければなりません。

BAMファイルの生成にfastq-listインプットを使用すると、RGID値1つにつき1つのリードグループが生成されます。BAMヘッダーには、以下のリードグループのRGタグが含まれます：

- ID（RGIDから）
- SM（RGSMから）
- LB（RGLBから）

各リードグループに対してタグを追加指定するには、列のタイトルを追加します。列のタイトルは、RGで始まる大文字4文字でなければなりません。例えば、PU（プラットフォームユニット）タグを追加するには、RGPUという名前の列を追加して、この列に、各リードグループの値を指定します。複数の列に、同じタイトルをつけることはできません。

fastq-listファイルには、複数サンプルのファイルが含まれます。fastq-listファイルにRGSMエントリーが1つしかない場合、そのほかのオプションを追加で指定しなくても、fastq-listファイルにリストされているファイルはすべてDRAGENにより処理されます。fastq-listファイルにRGSMエントリーが複数ある場合、*--fastq-list <filename>*に加えて、以下のいずれかを指定する必要があります。

- CSVファイルにある特定のサンプルを処理するには、`--fastq-list-sample-id <SampleID>`を使用します。fastq-listファイル内にRGSM値を持ち、指定されたSampleIDと一致するエントリーだけが処理されます。
- RGSM値に関係なく、1回のランで、すべてのサンプルをまとめて処理するには、`--fastq- list-all-samples`をtrueに設定します。

すべてのインプットリードグループは、同じサンプルに属すると想定されているので、1回のランで作成されるBAM出力ファイルとVCF出力ファイルは1つだけです。1回のBCL変換ランで、複数のサンプルを処理するには、`--fastq-list- sample-id`オプションの値を変えながら、DRAGENを複数回実行します。

より複雑なフィルターのために、RGSM値のグループ化やサブセットを指定するためのオプションはありませんが、fastq-listファイルを変更して、同じ効果を得ることができます。

以下に、必要な列を含むFASTQリストCSVファイルの例を示します：

```
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
```

体細胞インプットに`--tumor-fastq-list`オプションを使用するには、以下の例のように、`--tumor-fastq-list- sample-id <SampleID>`オプションを使って、対応するFASTQリストのサンプルIDを指定します：

```
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>
```

Tumor-Normalペアインプット

体細胞モードで、複数のサンプルを含む`fastq_lists`または`tumor_fastq_lists`（RGSM）を使用する場合、ループを使用して、2つのリストを順次処理し、テスト用のTumor-Normalペアを作成することができます。テスト対象の正常サンプルのRGSMを1行につき1つ使って`*.txt`ファイルを作成し、その後、テスト対象の腫瘍サンプルのRGSMを使って`*.txt`ファイルを作ります。腫瘍サンプルのRGSMが、対応する正常サンプルのRGSMと同じ順序で並んでいること、最後のサンプルの後に空白行が1行あることを確認してください。

体細胞モードでテストを行うには、以下のサンプルスクリプトを使用できます。反復1回につき、腫瘍サンプルリストから1エントリーと、正常サンプルリストから1エントリーが（上から下に）取得され、DRAGENランのインプットとして、Tumor-Normalペアが作成されます。

```
#!/bin/bash
```

```
HT="/staging/HT/"
tumor_fastq_list="/staging/inputs/tumor_fastq_list.csv"
normal_fastq_list="/staging/inputs/normal_fastq_list.csv"
tumor_samples_list="/staging/inputs/tumor_samples_list.txt"
normal_samples_list="/staging/inputs/normal_samples_list.txt"
while read -u 3 -r tumor_RGSM && read -u 4 -r normal_RGSM; do
output_dir="/staging/results/${tumor_RGSM}_${normal_RGSM}"
mkdir -p ${output_dir}
dragen \
-r ${HT} \
--tumor-fastq-list ${tumor_fastq_list} \
--tumor-fastq-list-sample-id ${tumor_RGSM} \
--fastq-list ${normal_fastq_list} \
--fastq-list-sample-id ${normal_RGSM} \
--output-directory ${output_dir} \
--output-file-prefix ${tumor_RGSM}_${normal_RGSM}
done 3<${tumor_samples_list} 4<${normal_samples_list}
```

以下に、このスクリプトでインプットとして使用されたFASTQリストとサンプルリストの例を示します。

```
Sample fastq_list.csv:
      RGPL,RGID,RGSM,RGLB,Lane,Read1File,Read2File
      DRAGEN_RGPL,DRAGEN_RGID_N1.1,normal-1,ILLUMINA,1,/staging/inputs/normal-
      1_S1_L001_R1_001.fastq.gz,/staging/inputs/normal-1_S1_L001_R2_
      001.fastq.gz
      DRAGEN_RGPL,DRAGEN_RGID_N1.2,normal-1,ILLUMINA,2,/staging/inputs/normal-
      1_S1_L002_R1_001.fastq.gz,/staging/inputs/normal-1_S1_L002_R2_
      001.fastq.gz
      DRAGEN_RGPL,DRAGEN_RGID_N2.1,normal-2,ILLUMINA,1,/staging/inputs/normal-
      2_S1_L001_R1_001.fastq.gz,/staging/inputs/normal-2_S1_L001_R2_
      001.fastq.gz
      DRAGEN_RGPL,DRAGEN_RGID_N2.2,normal-2,ILLUMINA,2,/staging/inputs/normal-
      2_S1_L002_R1_001.fastq.gz,/staging/inputs/normal-2_S1_L002_R2_
      001.fastq.gz
      DRAGEN_RGPL,DRAGEN_RGID_N3.1,normal-3,ILLUMINA,1,/staging/inputs/normal-
      3_S1_L001_R1_001.fastq.gz,/staging/inputs/normal-3_S1_L001_R2_
      001.fastq.gz
      DRAGEN_RGPL,DRAGEN_RGID_N3.2,normal-3,ILLUMINA,2,/staging/inputs/normal-
      3_S1_L002_R1_001.fastq.gz,/staging/inputs/normal-3_S1_L002_R2_
      001.fastq.gz
```

```
Sample tumor_fastq_list.csv content:
      RGPL,RGID,RGSM,RGLB,Lane,Read1File,Read2File
      DRAGEN_RGPL,DRAGEN_RGID_T1.1,tumor-1,ILLUMINA,1,/staging/inputs/tumor-1_
      S1_L001_R1_001.fastq.gz,/staging/inputs/tumor-1_S1_L001_R2_001.fastq.gz
      DRAGEN_RGPL,DRAGEN_RGID_T1.2,tumor-1,ILLUMINA,2,/staging/inputs/tumor-1_
      S1_L002_R1_001.fastq.gz,/staging/inputs/tumor-1_S1_L002_R2_001.fastq.gz
      DRAGEN_RGPL,DRAGEN_RGID_T2.1,tumor-2,ILLUMINA,1,/staging/inputs/tumor-2_
      S1_L001_R1_001.fastq.gz,/staging/inputs/tumor-2_S1_L001_R2_001.fastq.gz
      DRAGEN_RGPL,DRAGEN_RGID_T2.2,tumor-2,ILLUMINA,2,/staging/inputs/tumor-2_
      S1_L002_R1_001.fastq.gz,/staging/inputs/tumor-2_S1_L002_R2_001.fastq.gz
      DRAGEN_RGPL,DRAGEN_RGID_T3.1,tumor-3,ILLUMINA,1,/staging/inputs/tumor-3_
      S1_L001_R1_001.fastq.gz,/staging/inputs/tumor-3_S1_L001_R2_001.fastq.gz
      DRAGEN_RGPL,DRAGEN_RGID_T3.2,tumor-3,ILLUMINA,2,/staging/inputs/tumor-3_
      S1_L002_R1_001.fastq.gz,/staging/inputs/tumor-3_S1_L002_R2_001.fastq.gz
```

```
Sample normal_samples_list
      normal-1
      normal-2
      normal-3
```

```
Sample tumor_samples_list content
      tumor-1
      tumor-2
      tumor-3
```

FASTQ ORA入力ファイル

ORAファイルには、その他のFASTQ入力ファイルの種類と同じオプションを使用できます。ORAファイルを使用する場合は、FASTQファイル名をORAファイル名で置き換え、`--ora-reference`を使用して、ORAリファレンスディレクトリを指定します。

ORAリファレンスファイルの詳細については、346 ページの 「DRAGEN ORA圧縮と展開」 を参照してください。

以下のコマンドは、マッチする2つのORA FASTQファイルのペアエンド（-1と-2オプション）を表します。

```
      dragen -r <REF_DIR> -1 <fastq.ora1> -2 <fastq.ora2> \
      --ora-reference <LENADATA_DIR> \
      --output-directory <OUT_DIR> --output-file-prefix <OUT_PREFIX> \
```

```
    --RGID <RGID> --RGSM <RGSM>
```

BAM入力ファイル

マッパー/アライナーへのインプットとして、BAMファイルを使用するには、*--enable-map-align*をtrueに設定します。このオプションをfalse（初期設定）のままにしておいた場合、BAMファイルは、バリアントコーラーへのインプットとして使用できます。

BAMファイルをインプットとして指定すると、DRAGENは、入力ファイルに含まれるあらゆるアライメント情報を無視し、すべてのリードに対して新しいアライメントをアウトプットします。入力ファイルにペアエンドリードが含まれる場合、入力データのソートを指定して、ペアがいっしょに処理されるようにすることが重要です。他のパイプラインでは、インプットデータセットをリード名順に再ソートする必要があります。DRAGENは、インプットリードをペアリングし、特定されたペアをマッパー/アライナーへ送ることにより、この操作の速度を大幅に加速します。この機能の有効/無効を切り替えるには、*--pair-by-name*オプションを使用します（初期設定はtrueです）。

- 以下のように、*(-b)*オプションと`--pair-by-name=false`オプションを使って、BAMファイルのシングルエンドインプットを指定します：

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

- BAMファイルのペアエンドインプットを指定するには、以下のように、*(-b)*オプションと`--pair-by-name=true`オプションを使います：

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

CRAM入力ファイル

CRAMファイルを、DRAGENマッパー/アライナーとバリアントコーラーへのインプットとして使用できます。CRAMインプットを使用したときに利用可能なDRAGEN機能は、BAMインプットの場合と同じです。

`--cram-reference`オプションは必要なくなります。CRAM圧縮機能と展開機能は、DRAGENリファレンスを使用します。

マッパー/アライナー、またはバリアントコーラーにCRAMインプットを提供するには、以下のオプションを使用します：

- *--cram-input*：CRAMファイルの名前とパス。
- *--cram-input*：例えば、1つのCRAMファイルへのペアエンドインプットに使用できます。さらに、*--pair-by-name*オプションをtrueに設定します。

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

BCL入力ファイル

BCLは、イルミナシーケンスシステムのアウトプット形式です。一部の状況では、DRAGENは、マッピング/アライメント操作のためにBCLから直接読み込みを行い、FASTQへの変換にかかる時間を節約することができます。

DRAGENは、以下の状況で、BCLから直接読み込みできます：

- ランの一部としてインプットされるレーンが1つだけである（コマンドラインで指定）。
- そのレーンには、SampleSheet.csvファイルで指定されたサンプルが1つだけある。

BCLからFASTQへの変換が必要である場合、DRAGENは、BCL to FASTQ converterを提供します（307 ページの 「BCL変換」 を参照）。

以下に示すのは、インプットレーンが1つだけのBCLインプットのコマンド例です：

```
dragen --bcl-input-dir <BCL_ROOT> --bcl-only-lane <num> -r <ref_dir> \
--output-directory <out_dir> --output-file-prefix <out_prefix>
```

その他のBCL変換オプションについては、53 ページの 「入力ファイルの種類」 を参照してください。

N塩基の処理

シーケンス処理の最適化にDRAGENが使用する手法の1つが、Nベースコールに割り当てられた、塩基のクオリティスコアの上書きにつながる可能性があります。

`--fastq-n-quality`オプションと`--fastq-offset`オプションを使用すると、塩基のクオリティスコアが、固定塩基クオリティ値で上書きされます。これらのオプションの初期設定値は2と33で、これらは、イルミナの最小クオリティ35（ASCII文字の「#」）と一致します。

ペアエンドリードのリード名

一般的な慣例により、リード名には`/1`や`/2`のような接尾辞が含まれます。これらは、リードが表すペアの末尾を示します。`--pair-by-name option`を使用したBAMインプットについては、マッチするペア名の検索の際、これらの接尾辞は無視されます。初期設定では、DRAGENは、接尾辞の区切り文字としてスラッシュ（/）を使用し、名前を比較するときには`/1`や`/2`を無視します。初期設定では、DRAGENは、元のリード名から接尾辞を取り除きます。

DRAGENでは、接尾辞の使用方法を制御するために、以下のオプションが用意されています。

- 接尾辞の区切り文字を変更するには、`--pair-suffix-delimiter`オプションを使用します。このオプションにはスラッシュ（/）、ピリオド（.）、コロン（:）を指定できます。
- 名前全体を保護するには、`--strip-input-qname-suffixes`をfalseに設定します。
- すべてのリード名に新たな接尾辞を付加するには、`--append-read-index-to-name`をtrueに設定します。ここで使用される区切り文字は、`--pair-suffix-delimiter`オプションによって決まります。区切り文字の初期設定は、スラッシュです。したがって、名前には`/1`や`/2`が追加されます。

遺伝子アノテーション入力ファイル

RNA-Seqデータを処理するときに、*--annotation-file*オプションを使用して、遺伝子アノテーションファイルを提供することができます。このファイルを提供することにより、マッピングおよびアライメント段階の精度が向上します（271 ページの 「入力ファイル」 を参照）。このファイルは、GTF/GFF形式仕様に準拠していなければなりません。また、マッピングの対象となるリファレンスゲノムと一致するアノテーション付き転写産物を列挙している必要があります。GFF3形式も似ていますが、現時点ではサポートされていません。

DRAGENは、`SJ.out.tab`ファイル（272 ページの 「SJ.out.tab」 参照）をアノテーションファイルとしてとり、two-passモードの操作でアライナーのガイドを支援することができます。

ストリーム入力ファイル

DRAGENはAWS S3バケットから直接、または、HTTP署名付きURLを使用して、入力ファイルをストリームできます。処理の前に、入力ファイルをローカルディスクにダウンロードしておく必要はありません。これらのファイルは、ネットワーク経由で直接、DRAGENプロセッサーへストリームされます。

インプットストリームは、大きな入力ファイルにとって非常に有益です。DRAGENは、BAMとFASTQ圧縮ファイルのインプットストリームをサポートします。FASTQファイルの場合、インプットストリームは、シングルエンドFASTQ、ペアエンドFASTQ、およびFASTQリストを使用するすべての構成で使用できます。

インプットストリームは、以下の使用事例でサポートされます。

- FASTQおよびBAMのマッピング/アライメント。
- BAMからの生殖細胞系列および体細胞のスモールバリアントコーリング（再マッピングなし）。

サイズが極めて小さいその他のファイルタイプについては、データ解析を実行する前に、ローカルにダウンロードします。

セキュリティと許可

入力ファイルをストリームするには、リモートファイルへのアクセス許可を得る必要があります。S3オブジェクトでは、AWS認証と認証情報が必要です。AWS認証は、例えば、IAMポリシーなどを使って、実行するインスタンスにあらかじめ設定しておく必要があります。HTTP URLには、ほとんどの場合、クエリ文字列が添付されていて、この文字列に、認証情報、または許可を与えるために必要なトークンが含まれています。

例

DRAGENを使って、入力ファイルを直接ストリームする方法の例を以下に示します。

S3を使用したFASTQインプットのストリーム

```
dragen -f
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 s3://s3-bucket-name/path/to/object_1.fastq.gz \
-2 s3://s3-bucket-name/path/to/object_2.fastq.gz \
--RGID object_ID \
```

```
--RGSM sample_name \
--output-directory /staging/examples/ \
--output-file-prefix streaming
```

HTTPを使用したFASTQインプットのストリーム

```
dragen -f
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-1 https://bucket-name.amazonaws.com/path/to/object_
1.fastq.gz?querystring \
-2 https://bucket-name.amazonaws.com/path/to/object_
2.fastq.gz?querystring \
--RGID object_ID \
--RGSM sample_name \
--output-directory /staging/examples/ \
--output-file-prefix streaming
```

S3を使用したBAMインプットのストリーム

```
dragen -f
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-b s3://s3-bucket-name/path/to/object_1.bam \
--output-directory /staging/examples/ \
--output-file-prefix streaming
```

HTTPを使用したBAMインプットのストリーム

```
dragen -f
-r /staging/human/reference/hg19/hg19.fa.k_21.f_16.m_149 \
-b https://bucket-name.amazonaws.com/path/to/object_1.bam?querystring \
--output-directory /staging/examples/ \
--output-file-prefix streaming
```

サンプルの性別

バリアントコーラーなど、下流コンポーネントで使用される性核型をコントロールするには、*--sample-sex* コマンドラインオプションを使用します。コマンドラインを使用して、サンプルの性核型インプットが指定されていない場合、性核型が自動的に判断されます。性核型インプットは、バリアントコーリングで使用できるように、リファレンス性核型へ変換されます。他のコンポーネントでは、性核型インプットがサポートされている可能性があります。使用しているコンポーネントに対応するセクションを参照してください。

--sample-sexオプションは、以下の値をサポートしています。値では、大文字小文字の区別は行われません。

- none：性核型インプットはありません。コンポーネントでは、初期設定のリファレンス性核型が使用されます。
- auto：性核型は、Ploidy Estimatorにより推定されます。CNVコールを使用した場合、別の性推定モジュールを使って、性核型が決定されます。DRAGENが性核型を推定できない場合、コンポーネントに性核型はインプットされません。その後の動作は、noneの場合と同じです。初期設定値はautoです。
- female：性核型インプットはXXです。
- male：性核型インプットはXYです。

以下のコマンドライン例では、性核型の指定に--sample-sexが使われています。

```
--sample-sex FEMALE
--sample-sex MALE
--sample-sex NONE
```

値がnone、female、またはmaleであっても、Ploidy Estimatorを実行して、アウトプットを得ることができますが、バリアントコーラーは、コマンドラインで指定された性核型とは異なる推定性核型を使用することはありません。

性核型インプットは、以下のように、異なるコンポーネントで使用できるよう、リファレンス性核型へ変換されます。--sample-sexの使用方法の詳細については、関連するコンポーネントのセクションを参照してください。

性核型インプット	CNVコーラー	ExpansionHunter	Ploidy Caller	スモールバリアントコーラー	SVコーラー
XX	XX	XX	XX	XX	XXYY
XY	XY	XY	XY	XY	XXYY
XXY	XY	XX	XY	XXYY	XXYY
XYY	XY	XY	XY	XXYY	XXYY
X0	XX	XY	XX	XXYY	XXYY
XXXY	XY	XX	XY	XXYY	XXYY
XXX	XX	XX	XX	XXYY	XXYY
なし	XX	XX	XX	XXYY	XXYY

BAMおよびCRAM出力ファイルに対し自動生成されるMD5SUM

BAMおよびCRAM出力ファイルのMD5SUMファイルは、自動的に生成されます。MD5SUMファイルの名前は出力ファイルと同じですが、末尾に.md5sum拡張子が付加されています（例：`whole_genome_run_123.bam.md5sum`）。MD5SUMファイルは、出力ファイルのmd5sumを含む1行のテキストファイルで、このmd5sumは、Linux md5sumコマンドの出力と正確に一致します。

MD5SUMは、出力ファイルの書き込み時に計算されるので、パフォーマンスには大きな影響を与えません（Linuxのmd5sumコマンドと比較した場合。こちらは30x BAMで数分かかります）。

構成ファイル

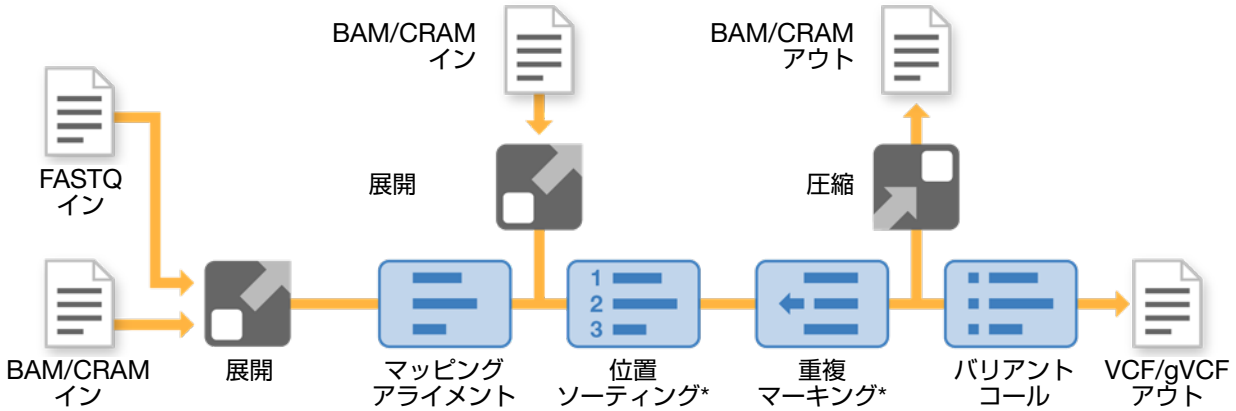
コマンドラインオプションは構成ファイルに保管できます。初期設定の構成ファイルの場所は、`/opt/edico/config/dragen-user-defaults.cfg`です。別のファイルを指定するには、`--config-file (-c)`オプションを使用します。ランに対して指定された構成ファイルは、そのランで使用される初期設定を提供します。この初期設定は、コマンドラインオプションを使ってオーバーライドできます。

さまざまな使用事例の初期設定を作成する場合には、`dragen-user-defaults.cfg`をテンプレートとして使用することを推奨します。構成ファイルは以下のように作成します。

1. `dragen-user-defaults.cfg`をコピーし、コピーの名前を変更します。
2. 使用事例に応じて、構成ファイルを編集します。
   オリジナルの`dragen-user-defaults.cfg`ではなく、コピーした方を編集してください。
   ランごとに変更されないオプションだけを追加することを推奨します。ランごとに変わるオプションを指定するには、コマンドラインを使用します。

DRAGEN DNA Pipeline

DRAGEN DNA PipelineはDRAGEN Bio-It Platformの強力な力を活用してNGSデータの二次解析を加速します。本パイプラインは、マッピング、スコアリング、ソーティング、重複マーキングおよびハプロタイプバリアントコーリングを行うための最適化された高度なアルゴリズムを搭載しています。



*オプション

DNAマッピング

DRAGENは主に、短いシードに完全に一致するリファレンスを見つけ出すことで、リードをマッピングします。ただし、DRAGENは単一のSNPで編集されたシードを検索することによっても、1ヌクレオチドごとにリファレンスと異なるシードをマッピングすることもできます。リードが長くなると少なくとも1つは完全マッチするシードを含む確率が高くなるため、シード編集は通常、ロングリード（100 bp以上）には必要ありません。また、ペアエンドを使用している場合、どちらかのメイトに一致するシードはそのペアのアライメントに成功するため、シード編集は必要ありません。ただし、シード編集は、マッピング時間が長くなるコストを伴いますが、シングルエンドのショートリードに対してマッピングの精度を向上させるために役立つことがあります。次のオプションでシード編集をコントロールします：

表 3　シード編集オプション

コマンドラインオプション名	構成ファイルオプション名
`--Mappper.seed-density`	seed-density
`--Mapper.edit-mode`	edit-mode
`--Mapper.edit-seed-num`	edit-seed-num
`--Mapper.edit-read-len`	edit-read-len

コマンドラインオプション名	構成ファイルオプション名
`--Mapper.edit-chain-limit`	edit-chain-limit

シード密度

`seed-density`オプションは、リファレンスゲノムへの完全マッチを見つけるために、マッパーが各リードからどれくらいの数の（通常重複している）一次シードをハッシュテーブル中で検索するかをコントロールします。

シード密度は0.0～1.0の間にしてください。内部で、DRAGENは要求された密度に等しいまたは近い使用可能なシードパターンを選択します。最も低密度のパターンは、32ポジションあたり1シード、すなわち、0.03125の密度です。最大密度値である1.0はリードのあらゆるポジションから開始するシードを生成します。

- **精度の検討**：一般的に、より高密度のシード検索パターンによりマッピング精度が向上します。ただし、ロングリード（50 bp以上）およびシーケンスエラー率が低い場合は、シード検索密度が初期設定の50%を超えると改善が最低限となります。
- **シード検討**：一般的に、より高密度のシード検索パターンによりマッピング速度が減速し、より低密度のシードパターンはマッピング速度を速めます。ただし、シードマッピング段階がアライメント段階よりも早く進んでいる場合、より低密度のシードパターンでもマッパーの速度は改善しません。

リファレンスシード間隔との関係

機能的に、より高密度またはより低密度のシード検索パターンは、`--ht-ref-seed-interval`オプションを介してリファレンスシード間隔をより短くまたはより長く指定するのと同様の影響があります。100%のリファレンスシードポジションを追加し50%のリードシードポジションを検索することは、50%のリファレンスシードポジションを追加し100%のリードシードポジションを検索することと同じ効果があります。いずれの方法でも、シードヒットの期待される密度は50%です。

もっと一般的に言うと、シードヒットの期待される密度はリファレンスシード密度とシード検索密度の積のことです。例えば、リファレンスシードの50%を追加し、リードシードポジションの33.3%（1/3）を検索する場合、期待されるシードヒット密度は16.7%（1/6）になります。

Local Analysis Softwareは自動的にシード検索パターンを調節し、リファレンスから追加されたシードポジションを系統的に見落とさないようにしています。例えば、偶数ポジションのみがハッシュテーブルに追加されている場合、リファレンスシード間隔が2であり、シード密度が0.5であっても、マッパーはリファレンスの奇数ポジションのみにマッチするシードを検索しません。

編集モードおよび鎖制限

`edit-mode`および`edit-chain-limit`オプションは、シード編集が使用されるタイミングをコントロールします。次の4つの`edit-mode`値が利用できます：

モードの値	説明
0	編集なし(初期設定)
1	鎖長テスト
2	ペアード鎖長テスト
3	シード全編集

編集モード0はすべてのシードが完全マッチするのに必要です。モード3は、リファレンスへの完全マッチが失敗したすべてのシードが編集されるため、最もコストが高くなります。

モード1および2はヒューリスティック（発見法）を用いて、精確なマッピングを行うために救済されるされた可能性が最も高いリードのみに対して編集されたシードを検索します。主なヒューリスティックは、シード鎖長テストです。完全に一致したシードは特定のリードに関して1番目のパスでリファレンスにマップされます。マッチするシードは同様にアライメントしたシードの鎖に分類されます。リード中の最長のシード鎖が閾値である`edit-chain-limit`を超える場合、既にマッピングしているポジションがあるため、そのリードはシード編集の必要はありません。

編集モード1は、シード鎖長テストを使って特定のリードに対するシード編集を開始します。シード鎖が`edit-chain-limit`を超えない、または完全にマッチするシードがない場合、編集されたシードを使って、2番目のシードマッピングパスを試みます。

編集モード2はペアエンドリードに対するヒューリスティックを最適化します。どちらかのメイトが`edit-chain-limit`よりも長い完全なシード鎖がある場合、もう片方のリードのシードマッチに基づいて、レスキュースキャンがそのメイトのアライメントを救済する可能性があるため、そのペアに対するシード編集は無効になります。編集モード2はシングルエンドリードに対してはモード1と同じです。

シード数およびリード長

編集モード1および2について、ヒューリスティックによりシード編集が開始されると、`edit-seed-num`および`edit-read-len`オプションはそのリードに対する2番目のパスでシードポジションを編集する数をコントロールします。完全マッチしたシードマッピングは、リードポジションの50%または100%から開始するシードのような、密集して重複しているシードパターンを使用できますが、シード編集の価値のほとんどは、より低密度のシードパターン、重複していないパターンですら編集することで得ることができます。一般的に、ユーザーアプリケーションがシード編集時のマッピング時間の追加を許容できるのであれば、多くのリードに対して低密度パターンでシードを編集することにより、同じ時間コストでマッピング精度を高めることができます。

シード編集が引き起こされるたびに、これらの2つのオプションは、リードの一番目の`edit-read-len`塩基に対して均一に分布する、`edit-seed-num`シード編集ポジションを要求します。`edit-seed-num`が6、`edit-read-len`が100である21塩基のシードの例では、編集された5'から連続するシードは、5塩基ずつ重複しずれて{0, 16, 32, 48, 64, 80}開始します。特定のリードが`edit-read-len`よりも短い場合、編集されるシードは少なくなります。

シード編集は--ht-ref-seed-intervalオプションが1を超える場合、コストがさらに高くなります。編集モード1および2では、追加のシード編集ポジションは自動的に生成され、追加したリファレンスシードポジションの見落としを回避します。編集モード3では、追加していないリファレンスポジションにマッチするクエリシードを一般的に見逃し、編集を引き起こすため、時間コストが劇的に増加することがあります。

マップの方向

--Mapper.map-orientationsオプションはバイサルファイトメチル化解析でリードをマッピングする際に使用します。このオプションは--methylation-protocolに対する設定値に基づいて自動的に設定されます。

--Mapper.map-orientationsオプションは、リードマッピングの方向をリファレンスゲノムのフォワード方向のみ、または逆相補方向のみに制限できます。次の値は、--map-orientationsに対する有効な値です：

- 0はどちらの方向も可能（初期設定）。
- 1はフォワード方向のマッピングのみ。
- 2は逆相補方向のマッピングのみ。

マッピング方向が制限されており、ペアエンドリードが使用されている場合、期待されるペア方向はFFまたはRFではなく、FRのみ可能です。

DNAアライメント

Smith-Watermanアライメントスコアリング設定

マッピングの第一段階では、リードからシードを生成し、リファレンスゲノムへの完全マッチを検索します。次に、シードマッチ結果は、シードがマッチしている最高密度の場所で完全なSmith-Watermanアライメントを実行することによって絞り込まれます。アライメントアルゴリズムは、リードの各ポジションをリファレンスのすべての候補ポジションと比較します。これらの比較は、リードとリファレンスとの可能性のあるアライメントのマトリクスに相当します。各候補アライメントポジションでは、スコアリングマトリクスを通過する間にSmith-Watermanアルゴリズムがスコアを生成します。このスコアは、アライメントがヌクレオチドマッチまたはミスマッチ、欠失、挿入に達したかどうかを反映します。リードとリファレンスのマッチはボーナスが与えられ、ミスマッチまたはIndelではペナルティが課されます。選択されたアライメントはマトリクスの中で全体的に最も高いスコアリングパスを有します。

複数の解釈が可能なアライメントでは、スコアに対して選択された特異的な値は、1つ以上のSNPとは対照的なIndelの可能性、またはクリッピングなしのアライメントを選択するバランスの取り方を示します。初期設定のDRAGENスコアリング値は、バリアントコールアプリケーションが、ヒトの全リファレンスゲノムに対して適度な長さのリードをアライメントするための合理的な値です。ただし、一連のSmith-Watermanスコアリングパラメーターは、ゲノム変異およびシーエンスエラーに関して曖昧なモデルを示します。別に調節したアライメントスコアリング値が一部のアプリケーションには適している場合があります。

次のオプションはSmith-Watermanアライメントをコントロールします：

オプション	説明
--Aligner.global	globalオプションはアライメントをリードのエンドツーエンドにするかをコントロールします。使用可能な値は0または1です。 • 1に設定したとき、グローバルアライメントのNeedleman-Wunschアルゴリズムに従って、アライメントは常にエンドツーエンドで行われます。アライメントはリファレンスにはエンドツーエンドで行われません。アライメントスコアは正または負の値となることがあります。 • 0に設定したとき、ローカルアライメントのSmith-Watermanアルゴリズムに従って、アライメントはリードのどちらかの末端または両端でクリップされることがあり、アライメントスコアは正の値となります。 ロングリードでは、値0が好ましいため、切断後の重要なリード断片はアライメントスコアを大幅に減少させることなくクリップされることがあります。切断の例には、大規模なIndel、構造多型、キメラリードなどがあります。リードの末端または末端付近での挿入は偽クリッピングとして機能することがあるため、オプションを1に設定することでロングリードに期待通りの効果をもたらさない場合があります。また、globalが0のとき、リードの様々な部分が大きく離れたリファレンスポジションにマッチする場合、複数の(キメラ)アライメントがレポートされることがあります。 ショートリードでは、globalを1に設定することが好ましい場合があります。ショートリードは構造的な切断が重複しておこる可能性が低く、キメラアライメントに対応できず、エンドツーエンドでうまくアライメントできない場合は、不正確なマッピングが疑われます。 クリップされていないアライメントに対して柔軟に優先性を持たせるようにするには、unclip-scoreオプションを使用するか、globalを1に設定せずにこの値を増すことを検討してください。
--Aligner.match-score	match-scoreオプションはリファレンスヌクレオチド(A、C、G、T)にマッチする、またはリファレンスの2～3のヌクレオチドのIUPAC-IUBコードにマッチするリードヌクレオチドのスコアを指定します。この値は0～15までの符号なし整数です。globalが1のときにのみ、match_scoreオプションを0に設定してください。一致スコアを高くすると長いアライメントが生じ、長い挿入は少なくなります。
--Aligner.match-n-score	match-n-scoreオプションは、リードポジションやリファレンスポジションがNコードである場合にアライメントされたポジションに対するスコアを指定します。このオプションは-16～15の符号付き整数です。

オプション	説明
--Aligner.mismatch-pen	mismatch-penオプションは、リードヌクレオチドがミスマッチしているリファレンスヌクレオチドまたはNを除くIUPAC-IUBコードに対するペナルティ、すなわち負のスコアを設定します。このオプションは0～63の符号なし整数です。ミスマッチペナルティが高いほど、SNPを回避するために挿入、欠失、クリッピングが多いアライメントとなります。
--Aligner.gap-open-pen	gap-open-penオプションは、ギャップ（挿入または欠失）を開けるためのペナルティ、すなわち負のスコアを設定します。値はギャップ0-base時にのみ当てはまります。ペナルティは常にgap-ext-penで乗じたギャップ長に可算されます。このオプションは0～127の符号なし整数です。ギャップオープンペナルティが高いほどアライメントCIGARにおいて、あらゆる長さの挿入および欠失が少なくなります。その代わり、SNPからのクリッピングまたはアライメントが使用されます。
--Aligner.gap-ext-pen	gap-ext-penオプションは、1塩基ごとにギャップ（挿入または欠失）を延長するためのペナルティ、すなわち負のスコアを設定します。このオプションは0～15の符号なし整数です。ギャップ延長ペナルティが高いほどアライメントCIGARの長い挿入と欠損が少なくなります。その代わり、短いIndel、クリッピングまたはアライメント上のSNPが使用されます。
--Aligner.unclip-score	unclip-scoreオプションは、リードの始点または終点に到達したアライメントに対するボーナススコアを設定します。クリップされていないときのボーナスが高いほど、過剰なSNPまたはIndelのないアライメントを実現し、アライメントがより頻繁にリードの始点や終点に到達します。エンドツーエンドのアライメントはボーナスを2回受け取ります。このオプションは0～127の符号なし整数です。 クリップされていないアライメントに対して柔軟に優先性を持たせるようにするために、globalが0のとき、0以外のunclip-scoreが役立ちます。globalが1のとき、両端のアライメントが強制されているため、クリップされていないときのボーナスはアライメントにほとんど影響を及ぼしません。ただし、no-unclip-scoreが1ではない限り、2倍のunclip-scoreがすべてのアライメントスコアに追加されます。 より長いリードでは、gap-open-penよりも高いunclip-scoreを設定することで、globalが1の時に生じるような、偽クリッピングとして使用されるリードの終点または終点付近の挿入の原因となることがあります。

オプション	説明
--Aligner.no-unclip-score	no-unclip-scoreオプションは0または1に設定できます。初期設定は1です。 no-unclip-scoreが1に設定されているとき、アライメントに寄与するクリップされていないときのボーナス(unclip-score)は、次のプロセスの前にアライメントスコアから除かれます。クリップされていないときのボーナスはaln-min-scoreとの比較、その他のアライメントスコアとの比較、ASまたはXSタグのレポート作成を含めることができます。ただし、クリップされていないときのボーナスは、特定のリファレンスセグメントに対してSmith-Watermanのアライメントによって見つけられたベストスコアのアライメントにも影響を及ぼすため、クリップされていないアライメントに対して偏りが生じます。 0を超えるunclip-scoreによって、Smith-Watermanのローカルアライメントに、リードの片方の末端または両端を延長させる場合、次のスコア変更が可能です: • no-unclip-scoreが0の場合、アライメントスコアは変わらないか増加します。 • no-unclip-scoreが1の場合、アライメントスコアは変わらないか減少します。 初期設定のno-unclip-scoreは1であり、globalが1の時はすべてのアライメントはエンドツーエンドで行われるため、この初期設定値が推奨されます。すべてのアライメントに同一のボーナスを追加する必要はありません。 no-unclip-scoreを変更する際は、aln-min-scoreを調節するかを検討してください。no-unclip-scoreが0の時、クリップされていないときのボーナスはaln-min-scoreの最低値と比較してアライメントスコアに含まれるため、aln-min-scoreによってフィルタリングで除外されるアライメントサブセットはno-unclip-scoreによって著しく変化することがあります。

オプション	説明
--Aligner.aln-min-score	aln-min-scoreオプションは最小の許容可能なアライメントスコアを指定します。低いスコアのアライメント結果は破棄されます。aln-min-scoreを増加または減少させることにより、マップされたリードの割合を減少または増加できます。このオプションは符号付き整数です。負のアライメントスコアはglobalが0の時に可能です。 aln-min-scoreオプションもMAPQの推定値に影響します。MAPQ算出の主要な寄与因子はベストなアライメントスコアと2番目にベストなアライメントスコアの差です。aln-min-scoreオプションは、ベストスコア以外に高いスコアが見つからなかった場合、最適ではないアライメントスコアとして機能します。従って、aln-min-scoreを増加することで、一部の低スコアリングアライメントに対してレポートされるMAPQが低下する場合があります。 min-score-coeffオプションを使用してリード長に応じてaln-min-scoreを調節できます。
--Aligner.min-score-coeff	min-score-coeffオプションはリード塩基ごとのaln-min-scoreの調節を行います。min-score-coeffオプションとaln-min-scoreオプションを同時に使用すると、リード長のアフィン関数として各リードに対する最小のアライメントスコアを定義できます。N塩基リードに対する最少スコアは次のように計算されます。 (min-score-coeff) *N+ (aln-min-score) min-score-coeffオプションは–64～63.999の範囲の整数です。値が0の場合、最小アライメントスコアはすべてのリード長に対するaln-min-scoreで固定されます。min-score-coeffに対して正の値を使うことで、ショートリードを低いアライメントスコアに一致させることができますが、ロングリードでは高いスコアを達成する必要があります。

ペアエンドオプション

DRAGENは一対のFASTQファイルから、または単一の交互的FASTQファイル中のパスしたペアエンドデータを処理できます。ハードウェアは2つの末端を別々にマップし、予測される方向でペアを形成する可能性が最も高いようであり、予測される大まかなインサートサイズを持つアライメントのセットを決定します。2つの末端に対するアライメントは、ペアリングのクオリティが評価され、予測されたサイズと離れたインサートサイズには高いペナルティが付きます。次のオプションはペアエンドデータの処理方法をコントロールします。

オプション	説明
--Aligner.pe-orientation	pe-orientationオプションは期待されるペアエンドの方向を指定します。方向性のあるペアのみが適切なペアとしてフラグ付けされます。次の値が有効です: • 0はFR（初期設定） • 1はRF • 2はFF
--Aligner.unpaired-pen	ペアエンドリードでは、ベストなマッピングポジションは各ペアに対して一緒に決定されます。各ポジションは、各メイトに対するアライメントのさまざまな組み合わせを検討し、見つかった大きなペアスコアに従って評価されます。ペアスコアは2つのアライメントスコアの合計から対合ペナルティを減じたもので、このアライメントペアよりも平均インサートが大きく異なるインサート長になりえないことを推定します。 unpaired-penオプションは、2つのアライメントが適切なペアポジションにない、または適切な方向ではないときに、どれくらいのアライメントペアスコアがペナルティを受けることになるかを指定します。また、このオプションは極端なインサート長のある適切なペアアライメントに対する最大の対合ペナルティとしても機能します。 unpaired-penオプションはMAPQに影響を及ぼす可能性があることから、Phredスケールで指定されています。内部で、このペナルティはSmith-Watermanのスコアリングパラメーターに基づいてアライメントスコアのマトリクスにスケール調整されます。
--Aligner.pe-max-penalty	pe-max-penaltyオプションは、あるリードのメイトが近くでアライメントしているため、そのリードについて推測したMAPQをどの程度増加できるかを制限します。ペアアライメントは、シングルエンドマッピングから取得したであろうMAPQ足すpe-max-penaltyの値よりも高いMAPQが割り当てられることは絶対にありません。 初期設定では、pe-max-penalty = mapq-max = 255であり、実質的にこの制限を無効にしています。 unpaired-penとpe-max- penaltyの主な違いは、unpaired-penは算出されたペアスコアに影響するため、どのアライメントが選択されるかに影響します。pe-max-penaltyオプションはペアアライメントに対してレポートされたMAPQのみに影響します。

平均インサートサイズの検出

ペアエンドデータを解析中の場合、DRAGENは2つの末端に対して最もクオリティが高いアライメントから可能性が考えられるペアを選択します。この選択を行うために、DRAGENはGaussian統計モデルを使用し、アライメントペアがペアを構成する尤度を評価します。このモデルは、特定のライブラリー調製ではおおよそ同じサイズの断片を生成する傾向があり、平均インサート長付近にインサート長が固まって分布するペアを産出するという洞察に基づいています。

入力ファイルとシングルリードグループからなるファイルについてライブラリー調製の統計の知識がある場合、インサート長分布に関する次の特性を指定できます：

- 平均値
- 標準偏差
- 四分位数

これらの特性は、pe-stat-mean-insert、pe-stat-stddev-insert、pe-stat-quartiles-insertおよびpe-stat-mean-read-lenオプションで指定できます。ただし、通常は、DRAGENがこれらの特性を自動的に検出できるようにすることが好まれます。

インサート長分布の自動サンプリングを有効にするには、--enable-samplingをtrueに設定します。ソフトウェアが実行を開始すると、アライナーを介して最大100,000ペアのサンプルのランを行い、分布を算出し、次にその結果として生じた統計を使用してインプットセット中の全てのペアを評価します。

DRAGENホストソフトウェアは次のとおりstdoutログにその統計をレポートします。

```
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)
```

各サンプルに対するインサート長分布はfragment_length_hist.csvに書き出されます。各サンプルは次のラインから開始します：

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

ラインの後はヒストグラムが続きます。

サンプルペアの数が非常に小さい場合、分布を高い信頼度によって特徴付ける十分な情報がありません。この場合DRAGENは、非常に広いインサート分布を指定する初期設定の統計値を適用します。これにより、アライメントペアが数万塩基離れて存在している場合であっても、適切なペアとしてアライメントのペアを認定する傾向があります。この場合、DRAGENは次のようなメッセージを出力します。

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

少数サンプルの式は次のように標準偏差を計算します：

```
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
```

初期設定モデルは標準偏差= 10000です。初めの100,000リードがマップされていないまたはすべてのペアが不正確なペアの場合、標準偏差は10,000に設定され、平均値および四分位数は0に設定されます。標準偏差の最小値は12であり、これはサンプル数と独立しています。

RNA-Seqデータの場合、インサートサイズ分布はペアにイントロンを含むため、通常と異なります。DRAGENソフトウェアはカーネル密度推定量を用いてその分布を推定し、長いテールをサンプルに一致させます。この推定はRNA-Seqデータと適切な対合のための精確な平均値と標準偏差につながります。

DRAGENは、`.insert-stats.tab`と呼ばれる出力ディレクトリにあるタブ区切りのログファイルに検出されたペアエンドの統計を書き出します。このファイルは、四分位数、平均値、標準偏差、最小値および最大値を含む各リードグループに対して検出されたインサートサイズの統計的分布を含みます。この情報は標準出力レポートと一致します。また、ログファイルはDRAGENがレスキュースキャンに適用したインサートの最小と最大の許容値を含みます。

レスキュースキャン

シードヒットが一方のメイトではなくもう片方のメイトに見つかったペアエンドリードでは、レスキュースキャンは平均インサート長のレスキュー範囲（rescue radius）内で不足しているメイトアライメントを探します。DRAGENの初期設定のレスキュー範囲は、経験的なインサート分布である標準偏差2.5に設定されています。インサート標準偏差がリード長よりも大きい場合、レスキュー範囲はマッピングの減速を制限するために限定されます。次のような警告メッセージが表示されます：

```
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
```

ユーザーはこの警告を無視するか、中間のレスキュー範囲を指定してマッピング速度を維持することができますが、マッピング感度を維持するにはレスキュー範囲に標準偏差2.5を使用することが推奨されます。レスキュースキャンを無効にするには、`max-rescues`を0に設定します。

出力オプション

DRAGENは各リードに対する複数の独立したアライメントを追跡できます。アライメントは、最適（一次）なアライメント、リードのさまざまなサブセグメントがマッピングしたアライメント（キメラ/補足）、およびリファレンスの異なる領域にリードが準最適に（二次）マッピングしたアライメントがあります。

初期設定のDNAアライメントでは、DRAGENは各リードに対して1つの一次アライメント、最大3つのキメラアライメント（supp-alignsは3）を出力できますが、二次アライメント（sec-alignsは0）は出力できません。supp-alignsまたはsec-alignsに対する最大のユーザー指定値は30です。リードあたり出力されるアライメントの最大合計数は31です。supp-alignsおよびsec-alignsの合計が30を超える場合、キメラアライメントが最優先で追跡されます。

次の構成オプションは、DRAGEN出力に含める各種アライメントの数をコントロールします。

オプション	説明
--Aligner.mapq-max	mapq-maxオプションは、あらゆるアライメントに対してレポートされる推定MAPQの制限値を指定します。0～255の値が有効です。算出されたMAPQが高い場合、その代わりにmapq-max値がレポートされます。初期設定は60です。
--Aligner.supp-aligns --Aligner.sec-aligns	supp-alignsおよびsec-alignsオプションは、各リードにレポートされる補足（例、キメラおよびSAM FLAG 0x800）アライメントおよび二次（例、準最適およびSAM FLAG 0x100）アライメントの最大値をそれぞれ制限します。 一次アライメント、補足アライメント、二次アライメントを含む、合計最大31のアライメントがリードに対しレポートされます。従って、supp-alignsおよびsec-alignsはそれぞれ0～30の範囲になります。 補足アライメントは二次アライメントよりも優先して追跡、出力されます。 これら2つのオプションを高値に設定すると、速度に影響を及ぼします。必要な分だけ増加してください。
--Aligner.sec-phred-delta	sec-phred-deltaオプションはレポートされた一次アライメントに比例したアライメントスコアに基づいて出力される二次アライメントをコントロールします。一次アライメントのPhred値範囲内にある二次アライメントのみがレポートされます。
--Aligner.sec-aligns-hard	sec-aligns-hardオプションは、出力できる数よりも多くの二次アライメントがある場合、すべての二次アライメントの出力を抑制します。すべての二次アライメントを出力できないとき、sec-aligns-hardを1に設定し、リードを強制的にマッピングされないようにします。
--Aligner.supp-as-sec	supp-as-secオプションが1に設定されている場合、補足（キメラ）アライメントはSAM FLAG 0x800ではなくSAM FLAG 0x100でレポートされます。初期設定値は0です。supp-as-secオプションはFLAG 0x800に対応しないツールとの互換性を提供します。

オプション	説明
`--Aligner.hard-clips`	`hard-clips`オプションは3ビットのフィールドとして使用し、0～7までの範囲の値を使用します。このビットは次のようにアライメントを指定します。 • ビット0は一次アライメント • ビット1は補足アライメント • ビット2は二次アライメント 各ビットは、その種類のローカルアライメントがハードクリッピング(1)またはソフトクリッピング(0)でレポートされるかを決定します。初期設定値は6です。これは一次アライメントはソフトクリッピングを使用し、補足アライメントおよび二次アライメントはハードクリッピングを使用することを意味します。

DRAGENグラフマッパー

DRAGENのグラフマッパーは、セグメントの重複およびイルミナリードでマップすることが困難なその他の領域におけるバリアントコール精度を改善します。グラフに基づくメソッドは、既知のアライメントを含むリファレンスに編み込まれている集団ハプロタイプに対してALT-awareマッピングを使用します。このメソッドでは、リードがシードマッピング可能でアライメント可能な別のグラフパスを設定します。グラフマッパーは、集団バリアントを含むリードが、それらのバリアントが観測された特異的領域に引き付けられるため、マッピングの曖昧さを低減します。

FASTAリファレンスをグラフリファレンスに展開するには、フェージングしたバリアントの集団ハプロタイプから由来する約900,000の短いオルタナティブコンティグを用いてDRAGENがFASTAリファレンスを拡張します。マッパーはALT-aware機能があり、この機能は、精確なリフトオーバーアライメントを用いて対応する一次アセンブリアライメントに集団ハプロタイプとマッチするリードを予測します。

一連の集団バリアント（VCF）またはハプロタイプが得られた場合、FASTAの変更は次の2種類に分類されます：

- **オルタナティブコンティグ**：この種類は集団ハプロタイプを表します。ALT-コンティグはシングルバリアントまたは近接するフェージングしたバリアントの組み合わせを含むことがあります。
- **曖昧（ambiguous）コード（IUPACコード）**：この種類はSNPを表します。アライメントを改善するには、分離した集団SNPを用いてリファレンスFASTAを編集します。

リードトリミング

DRAGENはハードウェアアクセラレーションによるリードトリミングを用いてリードからアーティファクトを除去できます。ハードウェアアクセラレーションによるリードトリミングは、DRAGENマッパーの一部としてU200およびAWSシステムで使用可能であり、追加のランタイムはかかりません。DRAGENは、さまざまな種類のアーティファクトまたは使用事例に向けた、複数の独立したトリミングフィルターを提供します。アーティファクトまたは使用事例を有効化して、構成し、目的の解析のためのリードトリミングをそれぞれ調整できます。リードトリミングは2種類の異なるモードである、ハードトリミングとソフトトリミングを使用します。

ハードトリミングモードを有効にするには、`--read-trimmers`を使用します。ハードトリミングモードでは、アーティファクトの可能性があるものはインプットリードから除去されます。20塩基未満にトリミングされるリードはフィルタリングされ、10N塩基を使用するプレースホルダーリードに置換されます。DRAGENはフィルタリングされるリードを0x200のフラグ付け設定に割り当てます。

DRAGENは新規の可逆圧縮ソフトトリミングモードを搭載しています。ソフトトリミングモードでは、リードはトリミングを介したものとしてマッピングされますが、除去される塩基はありません。ソフトモードでトリマーを有効にするには、`--soft-read-trimmers`を使用します。

ソフトトリミングは、アライメントされた出力中で実際にトリミングされた塩基を失うことなく、トリム可能なアーティファクトを含むリードの系統的なミスマッピングを抑制します。ソフトトリミングは、Poly-Gアーティファクトなどのトリム可能なアーティファクトを含むリードがリファレンスGホモポリマーにマッピングされるのを防ぐ、またはアダプターシーケンスがマッチするリファレンス座位にマッピングされるのを防ぎます。ソフトトリミングは、ソフトトリミングを使用していなければリードがマップされていると思われるところとは異なるリファレンスポジションにリードをマップする可能性があります。ソフトトリミングを使用中の場合、DRAGENはリードをフィルタリングせず、完全にトリミングされたと考えられる塩基を含むリードをマッピングしません。

Poly-Gアーティファクトに対するソフトトリミングはサポートされたシステムの初期設定で有効化されています。

リードトリミングツール

固定長トリミング

固定長トリミングは各リードの5’末端から固定した塩基数を除去します。固定されたサイズのアンプリコンからシーケンスデータを解析中であり、リード長が高クオリティのシーケンスデータの長さを一貫して超えることが予期される場合、固定長トリミングで予期される数を使用できます。

Poly-Gトリミング

Poly-Gアーティファクトは、合成が終了した後に色のない塩基Gがコールされる2色チャンネルのシーケンスシステムに現れます。結果的に、DRAGENは影響を受けたリードの末端にあるいくつかの誤った高い信頼度のあるG塩基をコールします。コンタミネーションが起こったサンプルでは、影響を受けた多くのリードが高いGコンテンツのあるリファレンス領域にマップされることがあります。影響を受けたリードは下流処理で問題を生じることがあります。

クオリティトリミング

塩基のクオリティは5’末端に向かうリードの長さとともに劣化し、合成の早期終了によるアーティファクトとは区別されます。低クオリティの塩基はマッピングおよびアライメント結果に影響することがあり、不正確な下流のバリアントコールまたはメチル化コールにつながる場合があります。クオリティトリミングツールは5’末端から内側に向かって継続的な平均塩基クオリティを算出し、最小数の塩基を除去するため、塩基の平均数は`--trim-min-quality`を用いて指定した閾値を上回ります。

アダプタートリミング

ライブラリー調製中の問題、すなわち小さな挿入のあるライブラリーが、使用したアダプター配列を含んでいる高クオリティなリードの合成をもたらすことがあります。解析前に除去されない場合、インサートがない塩基のマッピング効率と下流解析の精度を低下させることがあります。アダプタートリミングツールはFASTAファイルからアダプター配列を使用して、指定したサイズを超えるヒットすべてを除去します。アダプタートリミングは10%のミスマッチを許容します。

曖昧塩基トリミング

収量が低いまたはその他の制限により、クオリティトリミングが実行できない場合、別のオプションはリードの末端から明白な曖昧塩基のみを除去することです。有効な場合、曖昧塩基トリマーは、メイトペアの状況に関わらず、処理したリードすべての両末端に対してシンプルな完全マッチサーチを適用します。

最小長トリミング

上記のトリマーツールの実行後、各リードから固定塩基数を除去するために、最小長トリミングツールを使用して、トリマーの感度を最大にできます。例えば、各リードから5 bp除去したい場合、塩基5つが先に除去されるのであれば、7 bpのアダプターヒットが見落とされることがあります。この問題を軽減するため、DRAGENはオプションの最小トリム長フィルターを提供します。

最大長トリミング

小さなPCRアンプリコンなどの固定サイズのインサートのライブラリーを使用している場合、除去するための塩基数ではなく、すべてのリードを特定の長さにトリミングする長さを指定する方がより便利です。この場合には、最大長トリミングを使用することができます。

リードトリミングメトリクス

トリマーは`<output prefix>.trimmer_metrics.csv`という名前の付いたメトリクスを産生します。メトリクスはすべてのインプットデータの総計レベルで提供されます。このメトリクスユニットはリードまたは塩基の単位です。

メトリクス	説明
全インプットリード	入力ファイル中のリード総数。
全インプット塩基	インプットリード中の塩基総数。
全インプット塩基R1	R1リード中の塩基総数。
全インプット塩基R2	R2リード中の塩基総数。
平均インプットリード長	インプットリード数で割ったインプット塩基の総数。

メトリクス	説明
トリミングされた全リード	ソフトトリミングを含まない、少なくとも1つの塩基によってトリミングされたリード総数。
トリミングされた全塩基	ソフトトリミングを含まない、トリミングされた塩基総数。
リードあたりのトリミングされた平均塩基	インプットリード数で割ったトリミングされた塩基数。
トリミングされたリードあたりのトリミングされた平均塩基	トリミングされたリード数で割ったトリミングされた塩基数。
R1の3'末端に残存するpoly-G K-mer	トリミング後の推定Poly-Gアーティファクトを含むR1リードの3'末端数。
R2の3'末端に残存するpoly-G K-mer	トリミング後の推定Poly-Gアーティファクトを含むR2リードの3'末端数。
フィルタリングされた全リード	トリミング中にフィルタリングで除外されたリード数。
リードフィルタリングのためのR1の最小リード長	最小リード長を下回ってトリミングされているためにフィルタリングで除外されたR1リードの数。
リードフィルタリングのためのR2の最小リード長	最小リード長を下回ってトリミングされているためにフィルタリングで除外されたR2リードの数。
(トリマーツール)トリミングされたリード	トリマーによって少なくとも1塩基トリミングされたリード数。DRAGENはR1とR2の両メイトとトリミングされたリードのフィルタリング状況(フィルタリングされていないまたはフィルタリングされた)に対するメトリクスをレポートします。このメトリクスには、ソフトトリミング中にトリミングされたリードが含まれます。上記の各トリミングツールはメトリクスを生成します。
(トリマーツール)トリミングされた塩基	トリマーによってトリミングされた塩基数。このメトリクスはR1とR2の両メイトとトリミングされたリードのフィルタリング状況(フィルタリングされていないまたはフィルタリングされた)に対するメトリクスを生成します。このメトリクスには、ソフトトリミング中にトリミングされたリードからの塩基が含まれます。上記の各トリミングツールはメトリクスを生成します。

リードトリミング設定

次のオプションを使用してリードトリミングを構成します。

メトリクス	説明
--read-trimmers	ハードトリミングモードでトリミングフィルターを有効にするには、使用したいトリマーツールのカンマで区切られたリストに設定します。トリミングを無効にするには、noneに設定します。マッピング中にすべてのリードからアーティファクトが除去されます。リードトリミングは初期設定で無効にされています。有効なトリマーの名称は以下のとおりです: fixed-len：固定長トリミング polyg：Poly-Gトリミング quality：クオリティトリミング adapter：アダプタートリミング n：曖昧塩基トリミング min-len：最小長トリミング cut-end：最大長トリミング
--soft-read-trimmers	ソフトトリミングモードでトリミングフィルターを有効にするには、使用したいトリマーツールのカンマで区切られたリストに設定します。ソフトトリミングを無効にするには、noneに設定します。マッピング中に、トリミングされたかのようにリードはアライメントされますが、塩基はこれらのリードから除去されません。有効なトリマーの名称は以下のとおりです。ソフトトリミングは初期設定によってpolygフィルターに対して有効にされています。 fixed-len：固定長トリミング polyg：Poly-Gトリミング quality：クオリティトリミング adapter：アダプタートリミング n：曖昧塩基トリミング min-len：最小長トリミング cut-end：最大長トリミング
--trimming-only	リードトリミングのみを実行するためのマッピングとアライメントを無効にします。
--trim-r1-5prime	リード1の5'末端からトリミングされる塩基の固定数を指定します。
--trim-r1-3prime	リード1の3'末端からトリミングされる塩基の固定数を指定します。
--trim-r2-5prime	リード2の5'末端からトリミングされる塩基の固定数を指定します。
--trim-r2-3prime	リード2の3'末端からトリミングされる塩基の固定数を指定します。

メトリクス	説明
`--trim-min-length`	最小リード長を指定します。DRAGENは、リードトリミングが完了した後、最小リード長の値よりも短い長さのリードをフィルタリングします。
`--trim-min-quality`	リードの最小クオリティを指定します。DRAGENは最小クオリティ値を下回る塩基をリードの3'末端からトリミングします。
`--trim-adapter-read1`	アダプター配列を含むファイルを指定し、リード1の3'末端からトリミングします。
`--trim-adapter-read2`	アダプター配列を含むファイルを指定し、リード2の3'末端からトリミングします。
`--trim-adapter-stringency`	トリミングに必要なアダプター塩基の最小数を指定します。
`--trim-max-length`	両リードの配列からトリミングできる塩基の最大数を指定します。

DRAGEN FastQC

DRAGEN FastQCモジュールは、ハイスループットシーケンスデータのクオリティコントロールに使用される共通のメトリクスを計算するためのツールです。このツールはBabraham InstituteのFastQCツールによって生成されたメトリクスをモデルにしています。

このメトリクスはすべてのDRAGENマップアライメントワークフローで自動的に生成され、追加のランタイムはなく、`<PREFIX>.fastqc_metrics.csv`と呼ばれるCSV形式のファイルに出力します。

サンプルQCにのみ興味がある、またはFastQC結果のみを得たい場合、DRAGEN は`fastqc_metrics.csv`ファイルを直接生成するためのモードを提供します。

初期設定では、DRAGEN FastQCとリードトリミングは標準的なシーケンスアライメントワークフローに対する前処理ステップとして実行されます。DNAアライメントが必要ではない場合、またはQC結果がより早く必要になる場合、ワークフローのマッピングおよびBAM出力の一部を無効にできます。ワークフローは主要なメトリクスファイルのみを出力し、約70%早くランを実行します。このオプションを使用するには、DRAGENコマンド後に`--fastqc-only=true`を入力します。

メトリクス粒度

メモリー制約により、すべてのメトリクスに対して1塩基レベルの解像度を保証することはできません。DRAGENは`--fastqc-granularity`を介したビニングに対するアルゴリズムソリューションを提供します。DRAGENは各サイズまたはポジションに基づくメトリクスに対してメモリーに256ビンを割り当てます。4～7を含む粒度値はビンサイズを決定するために使用できます。高い値はより大きな解像度に対して小さなビンを使用します。低い値はより大きなリード長に対して大きなビンを作り出すために使用できます。

粒度	1塩基レベルの解像度(bp)	150での解像度(bp)	推奨されるリード長(bp)
7(初期設定)	1〜255	1	<256
6	1〜128	2	≥256〜<507
5	1〜64	4	≥507〜<4031
4	1〜32	8	≥ 4031

アダプターおよびKmerシーケンスファイル

アダプターまたはその他のシーケンスコンテンツに対するメトリクスを含めるには、DRAGEN FastQCは望まれる配列をFASTA形式で提供される必要があります。このために、DRAGENは次のオプションを提供します：

- アダプター配列には、`--fastqc-adapter-file`を使用します。
- 関心のあるkmerの追加には、`--fastqc-kmer-file`を使用します。

`--fastqc-kmer-file`オプションの場合、期待したアダプター結果を変更することなく関心のある配列を加えることができます。

DRAGEN FastQCは合計16までのアダプターとkmer配列に対応できます。各配列は、最長12 bpの長さまで可能です。初期設定では、DRAGENは`/opt/edico/config/adapter_sequences.fasta`にあるアダプターファイルを使用します。このファイルは、Babraham Institute（v 0.11.10以降）のFastQCと同じく、次のアダプター配列を含みます。

- イルミナユニバーサルアダプター：`AGATCGGAAGAG`
- イルミナSmall RNA 3'アダプター：`TGGAATTCTCGG`
- イルミナSmall RNA 5'アダプター：`GATCGTCGGACT`
- Nexteraトランスポゼース配列：`CTGTCTCTTATA`

FastQCメトリクス出力

FastQCメトリクスは、`<PREFIX>.fastqc_metrics.csv.`と呼ばれるラン出力ディレクトリにCSVファイル形式で出力されます。

レポートされるメトリクスはメトリクスタイプごとに8つのセクションに整理されます。各セクションは、長さ、ポジションまたはその他の関連するカテゴリカル変数ごとに別々の行に分類されます。次のようなメトリクスタイプによってセクションが構成されています。

オプション	説明
Read Mean Quality	リード総数。各Phredスケールクオリティ平均値は整数に四捨五入されます。

オプション	説明
Positional Base Mean Quality	特定のヌクレオチドを含み、リードの特定の場所にある塩基のPhredスケールクオリティ平均値。特定のポジションまたは範囲のいずれかの場所は初めに記載されます。ヌクレオチドは2番目に記載され、A、C、G、またはTにすることができます。N、すなわち曖昧塩基はシステム初期設定値、通常QV2を取ると考えられます。
Positional Base Content	リード中の特定の場所にある各特定のヌクレオチドの塩基数。特定のポジションまたは範囲のいずれかの場所は初めに記載されます。A、C、G、T、Nいずれかのヌクレオチドが2番目に記載されます。
Read Lengths	観察された各リード長のあるリード総数。長さは、`--fastqc-granularity`を使用して指定した設定に応じて、特定のサイズまたは範囲のいずれかを取ります。
Read GC Content	0%～100%の間の各GCコンテンツパーセンテージのリード総数。
Read GC Content Quality	0%～100%の間の各GCコンテンツパーセンテージのリードに対する平均Phredスケールのリードクオリティ平均値。
Sequence Positions	インプットリードの特定の位置から開始し、アダプターまたはその他のkmer配列が発見された回数。配列は引用符で囲まれたメトリクス説明に初めに記載されます。場所は2番目に記載され、特定のポジションまたは範囲のいずれかを取ります。
Positional Quality	特定の場所および分布の特定の分位点にある塩基に対するPhredスケールクオリティ値。場所は初めに記載され、特定のポジションまたは範囲のいずれかを取ります。分位点は2番目に記載され、0～100の整数を取ります。

次の例は各セクションからの列を含みます。

セクション	メイト	メトリクス	値
リード平均クオリティ	リード1	Q38リード	965377
ポジション単位の塩基の平均クオリティ	リード1	リードポジション145～152 Tの平均クオリティ	34.49
ポジション単位の塩基の平均クオリティ	リード1	リードポジション150 Tの平均クオリティ	34.44
ポジション単位の塩基の平均クオリティ	リード1	リードポジション256+ Tの平均クオリティ	36.99
ポジション単位の塩基コンテンツ	リード1	リードポジション145～152 A塩基	113362306

セクション	メイト	メトリクス	値
ポジション単位の塩基コンテンツ	リード1	リードポジション150 A塩基	14300589
ポジション単位の塩基コンテンツ	リード1	リードポジション256+ A塩基	13249068
リード長	リード1	150 bpリード長	77304421
リード長	リード1	144～151 bpリード長	77304421
リード長	リード1	255 bp以上リード長	1000000
リードGCコンテンツ	リード1	50% GCリード	140878674373
リードGCコンテンツクオリティ	リード1	50% GCリード平均クオリティ	36.20
配列ポジション	リード1	'AGATCGGAAGAG' 137 bp開始	20
配列ポジション	リード1	'AGATCGGAAGAG' 137～144 bp開始	23
ポジションクオリティ	リード1	リードポジション150、50%分位点QV	37
ポジションクオリティ	リード1	リードポジション145～152、50%分位点QV	37

ALT-awareマッピング

ヒトリファレンスのGRCh38は、このリファレンスの以前バージョンよりも多くのオルタナティブハプロタイプ、すなわちALTコンティグを含みます。一般的に、マッピングするリファレンスに複雑な領域に対するALTコンティグを含めることでマッピングとバリアントコール特異性が向上します。そうしない場合、ALTコンティグにマッチする一方で、一次アセンブリの一致する場所に対するスコアが低いリードはミスアライメントされることがあります。ただし、多くのリードがALTコンティグおよび一次アセンブリの一致するポジションに等しくアライメントされるときは、特別な処理をしないマッピングではバリアントコール感度が低下します。DRAGEN ALT-awareマッピングは、ALTコンティグからの特異性の向上を維持しながら、感度の低下問題を軽減します。

ALT-awareマッピングは指定したALTリフトオーバーアライメントで構築されたハッシュテーブルが必要です。詳細については、14 ページの 「ALTコンティグハッシュテーブル」 を参照してください。リフトオーバーアライメントで構築したハッシュテーブルが提供されている場合、DRAGENは自動的にALT-awareマッピングでランを実施します。リフトオーバーリファレンスを用いてALT-awareマッピングを無効にするには、`--alt-aware`オプションをfalseに設定します。

ALTコンティグがリファレンスhg19またはGRCh38で検出される場合、DRAGENはALT-awareハッシュテーブルを必要とします。DRAGENでこの要件を無効にするには、`--ht-alt-aware-validate`オプションをfalseに設定します。

ALT-awareマッピングが有効のとき、DRAGENのマッパーとアライナーはALTコンティグポジションと一致する一次アセンブリポジションとのリフトオーバーの関係を認識します。DRAGENはALTコンティグ内のシードマッチを使って、後者のスコアが悪い場合であっても、一致する一次アセンブリアライメントを取得します。一次アセンブリアライメントの候補および同じ場所へと変換したゼロ個以上のALTアライメント候補をそれぞれ含む、リフトオーバーグループが形成されます。適切なペアアライメントを考慮しながら、各リフトオーバーグループはベストマッチするアライメントに従ってスコアリングされます。

最も高いスコアのリフトオーバーグループから、一次出力アライメントとして一次アセンブリの代表と、2番目にベストなリフトオーバーグループに対するスコアの差に基づいて算出されたMAPQが得られます。一次アセンブリ内に一次アライメントを出力することで正常なアライメントカバレッジを維持し、そこでのバリアントコールを容易にします。`--Aligner.en-alt-hap-aln`オプションを1に設定し、`--Aligner.supp-aligns`が0を超えている場合、一致するオルタナティブハプロタイプアライメントも出力され、補足アライメントとしてフラグ付けされることがあります。

以下は、オルタナティブハプロタイプを処理するための代替アプローチの比較です。

- ALTコンティグなしのリファレンスへのマッピング：
  - ALTコンティグにマッチするALTコンティグはミスアライメントする可能性があり、偽陽性のバリアントコールをもたらすことがあります。
  - ALTコンティグにマッチするALTコンティグが一次アセンブリと大幅に異なる場合、マッピングとバリアントコールの感度は低くなります。
- ALTコンティグはあるが、alt awareのないマッピング：
  - ALTコンティグにマッチするALTコンティグがミスアライメントせず、関連する偽陽性のバリアントコールが阻止されます。
  - 一部のリードがALTコンティグにマッピングしているため、オルタナティブハプロタイプによってカバーされている一次アセンブリ領域のアライメントカバレッジが低い、または0になります。
  - オルタナティブハプロタイプが一次アセンブリに類似しているまたは同一の場合、オルタナティブハプロタイプによってカバーされている領域にあるMAPQが低い、または0になります。
  - バリアントコール感度はオルタナティブハプロタイプによってカバーされている領域全体で低下します。
- ALTコンティグおよびalt awareのあるマッピング：
  - ALTコンティグにマッチするALTコンティグがミスアライメントせず、関連する偽陽性のバリアントコールが阻止されます。
  - 一次アライメントが一次アセンブリにマッピングするため、オルタナティブハプロタイプによってカバーされている領域の通常のアライメントカバレッジです。
  - リフトオーバーグループ内のアライメント候補は競合すると考えられないため、通常のMAPQが割り当てられます。
  - リードがマッチするALTコンティグが一次アセンブリと大幅に異なる場合、マッピングとバリアントコールは良好な感度になります。

ソーティング

マッピング/アライメントシステムは、初期設定ではリファレンスシーケンスとポジションごとにソートされたBAMファイルを生成します。このBAMファイルを作成することにより、`samtools sort`または同等のポストプロセスコマンドを実行する必要性を除去します。BAMファイルの作成を有効または無効にするために、次のように`--enable-sort`オプションを使用できます。

- 有効にするには、trueに設定します。
- 無効にするには、falseに設定します。

リファレンスハードウェアシステムでは、30xの全ゲノムでソートが有効化されたランではランタイムが約6～7分増加します。

重複バリアントのフィルタリング

DRAGENは別々のVCFファイルに共通するバリアントを見つけて除去できます。DRAGENは次のモードに対応します：

- **小さなIndel重複除去**：構造多型VCFとスモールバリアントVCFを使用中の場合、DRAGENはスモールバリアントVCFに現れる構造多型VCF中の小さなIndelすべてをフィルタリングします。バリアントをノーマライズするためにVCFファイルを生成するには、リファレンスゲノムを提供する必要があります。DRAGENは、最大500塩基ごとにトリミングおよび左シフトを行うことでバリアントをノーマライズします。
- **SMN重複除去**：スモールバリアントVCFとExpansionHunter VCFを使用中の場合、DRAGENはINFOタグVARID=SMNのあるExpansionHunter VCF中のラインと同じ染色体とポジションを有するスモールバリアントVCF中のラインをフィルタリングします。リファレンスゲノムは必要ありません。

VCFまたはgVCFファイルを入力するには、次のコマンドラインオプションを使用します。入力ファイルは変更されません。

- `vd-sv-vcf`：構造多型VCFまたはgVCFを指定します。
- `vd-small-variant-vcf`：スモールバリアントのVCFまたはgVCFを指定します。
- `vd-eh-vcf`：ExpansionHunter VCFまたはgVCFを指定します。

DRAGENは次のような出力ファイルの名称と種類を特定します。

コンポーネント	説明
出力プレフィックス	値が`output-file-prefix`に対して指定されている場合、そのプレフィックスは通常通り使用されます。 値が有効ではない場合、フィルタリングされたインプットの名前がプレフィックスとして使用されます。
重複除去モード	使用する重複除去モードに応じて、プレフィックスの後に`.small_indel_dedup`または`.smn_dedup`が続きます。
ファイルの種類	出力ファイルの種類は入力ファイルの種類（VCFまたはgVCF）に一致します。 `enable-vcf-compression`がtrueに設定されている場合、入力ファイルが圧縮されていたかに関わらず、出力ファイルはgzip形式で圧縮されます。

コマンドラインオプション

バリアント重複除去のために、次のコマンドラインオプションを使用します。

オプション	説明
`enable-variant-deduplication`	バリアント重複除去を有効にするには、`true`に設定します。 初期設定は`false`です。
`enable-vcf-indexing`	Tabixのインデックスファイルを生成するには、`true`に設定します。 初期設定は`true`です。

オプション	説明
vd-output-match-log	テキストファイルにマッチするラインを記録するには、trueに設定します。初期設定はfalseです。各マッチでは、二つのマッチするラインがそれぞれをフォローし、次に新しいラインが続きます。マッチログの名前は、使用する重複除去モードに応じて、match_log.smn_dedup.txtまたはmatch_log.small_indel_dedup.txtのどちらかになります。

以下はSMN重複除去の独立したランを行うためのコマンドの例です：

```
dragen --enable-map-align false \
--enable-variant-deduplication true \
--vd-small-variant-vcf <small variant vcf> \
--vd-eh-vcf <expansion hunter vcf> \
--output-directory /tmp/ \
--vd-output-match-log true \
```

また、構造多型とスモールバリアントコーラーの両方が有効の場合、DRAGENジョイントコーラーからの出力について、小さなIndelの重複除去も自動的に実行できます。小さなIndelの重複除去を自動で実行するには、enable-variant-deduplicationをtrueに設定し、vd-sv-vcf、vd-small-indel-vcfおよびvd-eh-vcfのインプットオプションが設定されていないことを確認してください。小さなIndelの重複除去のみが自動で実行されます。

以下は小さなIndel重複除去の自動ランを行うためのコマンドの例です。

```
dragen
--ref-dir <REF>
--output-directory <DIR> \
--output-file-prefix <PREFIX> \
-b <BAM>
--enable-map-align false \
--enable-variant-caller=true"
--enable-sv=true"
--enable-variant-deduplication=true"
--vd-output-match-log=true"
```

重複マーキング

重複アライメントリードのマーキングまたは除去は全ゲノムシーケンスに共通するベストプラクティスです。このマーキングまたは除去を行わない場合、バリアントコールにバイアスを生じ、不正確な結果につながります。

DRAGENシステムは重複リードをマークまたは除去でき、FLAGフィールドにマークされた重複または完全に除去された重複のどちらかを含むBAMファイルを生成します。

アルゴリズム

DRAGENの重複マーキングアルゴリズムはPicardツールキットのMarkDuplicates機能をモデルにしています。アライメントされたすべてのリードはサブセットに分類されます。各サブセットのすべてのリードは重複している可能性があります。

2つのペアが重複するには、次の要件を満たす必要があります：

- 両末端のアライメント座標が同一。CIGARより、ソフトクリップまたはハードクリップのためにポジションが調整されている。
- 2つの末端の方向からの向きが同一であり、最も左端の座標が一番目に来ている。

ペアではないリードは、他のリードのいずれかの末端に対して同一の座標および同一の向きがある場合、ペアであってもペアでなくても、重複とマークされることがあります。

マッピングされないまたはリードペアは重複としてマークされることはありません。

DRAGENは重複グループを同定し、そのグループのうちのベストな重複を選択し、その他をBAM PCRまたは光学系の重複フラグ（0x400または10進数1024）でマークします。比較のために、シーケンスの平均Phredクオリティに基づいて、重複をスコア化します。ペアは両端のスコアの合計を受け取りますが、ペアではないリードはマッピングした一端のスコアを得ます。このスコアは最も高いクオリティのベースコールのあるリードを保存するためのものです。

2つのペア（またはリード）がクオリティスコアに完全マッチしている場合、DRAGENは高い方のアライメントスコアのあるペアを選択して同点のスコアに差をつけます。アライメントスコアも同じ複数のペアがある場合では、DRAGENは任意にペアを選択します。

ペアではないリードRのスコアは、塩基ごとの平均Phredクオリティスコアとなり、次のよう計算されます：

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

ここで：

- RはBAMレコードとします。
- QUALはPhredクオリティスコアのアレイとします。
- `dedup-min-qual`は初期設定値が15のDRAGEN構成オプションです。

ペアの場合、スコアは2つの末端に対するスコアの合計です。

スコアは半角数字で保存され、値は4分の1に四捨五入されます。四捨五入により、Picardで選択された重複マーキングとは異なる重複マーキングが生じることがあります。ただし、これらのリードはクオリティが類似しているため、バリアントコールの結果に及ぼす影響は無視できる程度です。

制限

DRAGEN重複マーキングの実施には次の制限が当てはまります：

- Phredシーケンスクオリティスコアが近い2つの重複リードまたはペアがある場合、DRAGENはPicardが選択したものと異なる方を選択する場合があります。この違いによるバリアントコール結果に与える影響は無視できる程度です。
- インプットとしてFASTQファイル1つを使用している場合、DRAGENはコマンドライン引数（PGLB）として1つのライブラリーIDのみを受け入れます。このため、システムへのFASTQインプットはライブラリーIDごとに前もって別々にしておく必要があります。ライブラリーIDは重複していないことを区別するための基準として使用できません。

設定

以下はDRAGENで重複マーキングを構成するために使用できるオプションです：

オプション	説明
`--enable-duplicate-marking`	重複マーキングを有効にするには、trueに設定します。`--enable-duplicate-marking`が有効のとき、`enable-sort`オプションの値に関わらず、その出力はソーティングされます。
`--remove-duplicates`	重複レコードの出力を抑制するには、trueに設定します。falseに設定した場合、重複BAMレコードのFLAGフィールドで0x400のフラグ付けを設定します。`--remove-duplicates`が有効のとき、`enable-duplicate-marking`も同様に有効です。
`--dedup-min-qual`	重複リード中から選択を行うために、塩基に対する最小のPhredクオリティスコアを指定し、使用したクオリティスコアの算出値にこのスコアを含めます。

スモールバリアントコール

DRAGENスモールバリアントコーラーはハードウェアとソフトウェアのハイブリッドで実施する高速ハプロタイプコーラーです。このコーラーは、候補ハプロタイプを生成するために関心領域中でローカライズした*de novo*アセンブルを実施し、隠れマルコフモデル（HMM）を用いてリード尤度の計算を実施します。

バリアントコールは初期設定で無効にされています。バリアントコールを有効にするには、`--enable-variant-caller`オプションをtrueに設定します。

バリアントコーラーアルゴリズム

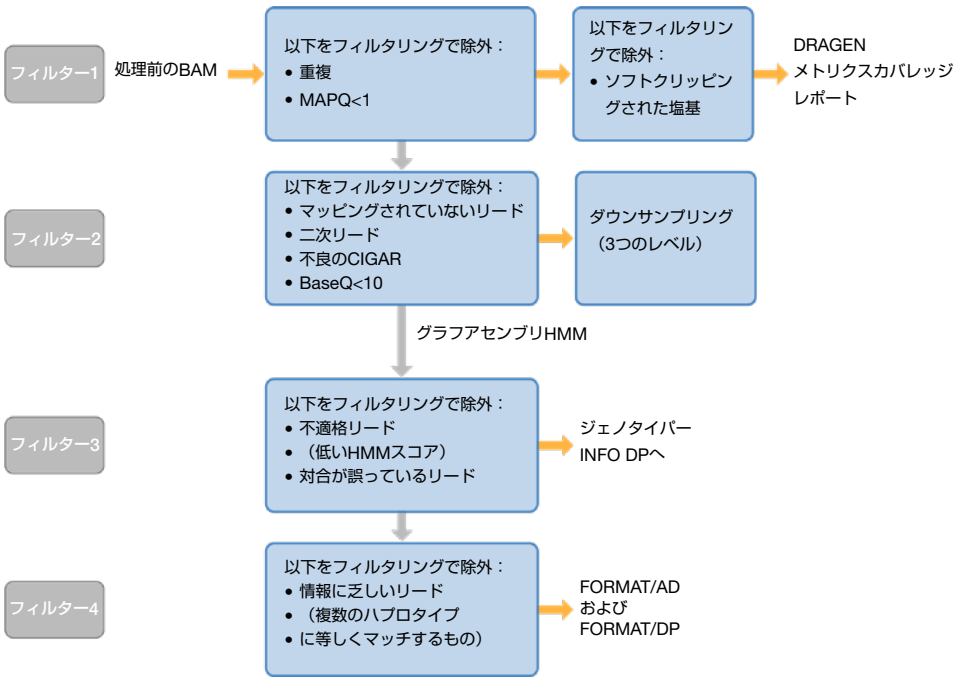
DRAGENスモールバリアントコーラーは次のステップを実施します：

- **活性領域の同定**：リファレンスと一致しない複数のリードが同定される領域を同定し、処理を行うためにそれら周辺の範囲（アクティブ領域）を選択します。
- **ローカライズしたハプロタイプアセンブリ**：各アクティブ領域では、de Bruijnグラフ（DBG）に各アクティブ領域の重複リードすべてをアセンブルします。DBGは各リードまたは複数のリードの重複するKmer（長さKの配列）に基づいた有向グラフです。すべてのリードが同定されたとき、DBGは線形になります。相違がある場合、このグラフは複数のパスが分岐し、再結合する領域を形成します。ローカル配列の繰り返しが多すぎ、Kが小さすぎる場合、グラフが無効とされる円形が形成されます。K値= 10および25は初期設定で試行されます。これらの値が無効なグラフを生成する場合、円形のないグラフが得られるまで、追加の値としてK=35、45、55、65が試行されます。この円形なしのDBGから、可能性のあるすべてのパスが抽出され、候補ハプロタイプの完全なリストを生成します。すなわち、真のDNA配列が少なくとも1本のストランドにあるかもしれないという仮定に基づいて行われます。

- **ローカライズしたハプロタイプアセンブリ**：de Bruijnグラフ（DBG）に各アクティブ領域の重複リードすべてをアセンブルします。DBGは各リードまたは複数のリードの重複するKmer（長さKの配列）に基づいた有向グラフです。すべてのリードが同定されたとき、DBGは線形になります。相違がある場合、このグラフは分岐し再結合する複数のパス領域を形成します。ローカル配列の繰り返しが多すぎ、Kが小さすぎる場合、このグラフを無効にして、サイクルを実行できます。DRAGENは初期設定値としてK=10および25を使用します。これらの値が無効なグラフを生成する場合、サイクルなしのグラフが得られるまで、追加の値としてK=35、45、55、65が試行されます。このサイクルなしのDBGから、DRAGENは可能性のあるすべてのパスを抽出し、候補ハプロタイプの完全なリストを生成します。すなわち、真のDNA配列が少なくとも1本のストランドにあるかもしれないという仮定に基づいて行われます。
- **ハプロタイプアライメント**：Smith-Watermanアルゴリズムを使用して、抽出した各ハプロタイプをリファレンスゲノムにアライメントします。このアライメントはリファレンスからのどのようなバリエーションが存在するかを特定します。
- **リード尤度計算**：特定のハプロタイプが真のオリジナルの検証したDNAであると仮定して、各ハプロタイプに対する各リードを検証しリードを観察する確度を推測します。この計算はペアの隠れマルコフモデル（HMM）を評価することで実施されます。これにより、PCRによって、または観察したリードにシーケンスエラーが生じることによってハプロタイプが変更されたかもしれないさまざまな可能性を考慮します。HMM評価は動的なプログラミングメソッドを使用して、観察されたリードに達したマルコフ状況の推移の合計確度を算出します。
- **ジェノタイピング**：候補ハプロタイプから可能性のある二倍体の変異の組み合わせを形成し、各組み合わせに対して、集積したリード全体を観測している条件付き確度を計算します。計算には、ペアHMM評価からの各ハプロタイプと仮定して、観測する各リードの構成確度を使用します。これらの計算はベイズの式に当てはめられ、集積したリード全体が観察されたと仮定して、各遺伝型が解析されているサンプルの遺伝型である尤度を計算します。最大尤度の遺伝型がレポートされます。

BAMインプットのフィルタリング

バリアントコール中に適用されたフィルタリングステップにより、IGV BAM集積の情報はVCF/gVCFのINFO/DPおよびFORMAT/DPとは異なります。4つのフィルターがあり、ジェノタイピング計算からリードを除外するために使用します。次の図に4つのフィルターの要約を示します。



- フィルター1はIGV BAMインプットから次のリードをフィルタリングして除外します：
  - 重複リード。
  - MAPQ=0のリード。
  - ソフトクリッピングされた塩基。DRAGENはカバレッジレポートを計算するときのみソフトクリッピングされた塩基をフィルタリングして除外します。
  - **（体細胞）**vc-min-tumor-read-qual >1のとき、MAPQ <vc-min-tumor-read-qualのリード。
- フィルター2はBQ <10の塩基をトリミングし、次のリードをフィルタリングして除外します：
  - マッピングされていないリード。
  - 二次リード。
  - CIGARが不良のリード。
- フィルター3はダウンサンプリングとHMM後に行います。フィルター3は次のリードをフィルタリングして除外します：
  - 対合が誤っているリード。対合が誤っているリードとは、ペアが2つの異なるリファレンスコンティグにマッピングされているリードです。
  - 不適格リード。リードのHMMスコアが閾値よりも低い場合、リードは不適格とされます。
- フィルター4はジェノタイパーの実行後に行います。ジェノタイパーはアノテーション情報をFORMATフィールドに追加します。フィルター4は情報に乏しいリードをフィルタリングして除外します。例えば、2つの異なるハプロタイプに対するリードのHMMスコアがほぼ等しい場合にそのリードがフィルタリングで除外される理由は、2つのハプロタイプのどちらがより可能性が高いかを区別するための十分な情報を提示していないためです。
  INFO/DPは情報が豊富なリードおよび情報に乏しいリードの両方を含みます。
  FORMAT/ADおよびFORMAT/DPは情報が豊富なリードのみを含みます。

バリアントコーラーオプション

次のオプションはバリアントコーラーステージをコントロールします。

オプション	説明
--enable-variant-caller	*--enable-variant-caller*をtrueに設定し、DRAGENパイプラインのバリアントコーラーステージを有効にします。

--vc-target-bed	**(オプション)**スモールバリアントコーラー、ターゲットBED関連カバレッジ、およびBEDファイルに指定された領域へのCallabilityメトリクスを制限します。BEDファイルは少なくともタブで区切られた3列を含むテキストファイルです。初めの3列は、染色体、開始ポジション、および終了ポジションです。ポジションはゼロベースの値です。例えば:
	`# header information`
	`chr11 0 246920`
	`chr11 255660 255661`
	バリアントのリファレンス距離がターゲットBEDのいずれかの領域と重複する場合、そのバリアントは出力されます。リファレンス距離が重複しない場合、そのバリアントは出力されません。SNPおよび挿入では、リファレンス距離は1 bpです。欠失では、リファレンス距離は欠失の長さです。

--vc-target-bed-padding	**(オプション)**指定した値ですべてのターゲットBED領域を広げます。例えば、BED領域が1:1000～2000であり、指定したパディング値が100の場合、結果は1:900～2100のBED領域とパディング値0を用いた結果と等しくなります。*--vc-target-bed-padding*に追加したパディング値は、スモールバリアントコーラーおよびターゲットBEDカバレッジ/コール可能性レポートに使用されます。初期設定パディング値は0です。
--vc-target-coverage	ダウンサンプリング用のターゲットカバレッジを指定します。初期設定値は、生殖細胞系列モードでは500、体細胞モードでは50です。
--vc-enable-gatk-acceleration	GATKモードで実行するには`true`に設定します。このオプションを使用したGATKモードの有効化は、生殖細胞系列モードのGATK 3.7および体細胞モードのGATK 4.0と一致します。
--vc-remove-all-soft-clips	`true`に設定し、ハプロタイプアセンブリステップ中にソフトクリッピングされた塩基を無視します。
--vc-decoy-contigs	カンマ区切りのリストで、バリアントコール中にスキップするコンティグを指定します。このオプションは構成ファイル中で設定できます。
--vc-enable-decoy-contigs	trueに設定し、デコイコンティグ上でのバリアントコールを有効にします。初期設定値はfalseです。

オプション	説明
--vc-enable-phasing	可能な場合バリアントのフェージングを有効にします。初期設定値はtrueです。
--vc-combine-phased- variants-distance	組み合わされるフェージングされたバリアント間の最大距離を設定します。初期設定値は0であり、オプションを無効にします。

スモールバリアントコールのダウンサンプリングオプション

スモールバリアントコールパイプラインでのリードのダウンサンプリングでは、以下のオプションが使用できます。

オプション	説明
--vc-target-coverage	開始位置が所定の位置と重複するリードの最大数を指定します。
--vc-max-reads-per-active-region	所定の活性領域をカバーするリードの最大数を指定します。
--vc-max-reads-per-raw-region	所定の処理前領域をカバーするリードの最大数を指定します。
--vc-min-reads-per-start-pos	開始位置が所定の位置と重複するリードの最小数を指定します。
--high-coverage-support-mode	trueに設定されている場合、高カバレッジモードのダウンサンプリングオプションを適用します。カバレッジが1000xを超えるターゲットパネルでは、このオプションを有効にすることを推奨しますが、ランタイムの速度が低下します。

ミトコンドリアのスモールバリアントコールでは、ダウンサンプリングオプションを個別に設定できます。これは、ミトコンドリアのコンティグが、WGSデータセットの残りのコンティグよりも深度が大きいためです。以下は、ミトコンドリアのコンティグのダウンサンプリングオプションです。

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

ターゲットカバレッジと処理前/活性領域オプションの最大/最小リード数は直接には関係しておらず、独立して実行することができます。

最初にターゲットカバレッジオプションを実行しますが、これは所定の位置で同じ開始位置を共有するリード数を制限することが目的です。これは、所定の位置での合計カバレッジの制限ではありません。

以下は、それぞれのスモールバリアントコールモードのダウンサンプリングの初期設定値です。

モード	ダウンサンプリングオプション	初期設定値
生殖細胞系列	*--vc-target-coverage*	500
生殖細胞系列	*--vc-max-reads-per-active-region*	10000
生殖細胞系列	*--vc-max-reads-per-raw-region*	30000

モード	ダウンサンプリングオプション	初期設定値
体細胞	*--vc-target-coverage*	50
体細胞	*--vc-max-reads-per-active-region*	10000
体細胞	*--vc-max-reads-per-raw-region*	30000
高カバレッジ	*--vc-target-coverage*	100000
高カバレッジ	*--vc-max-reads-per-active-region*	200000
高カバレッジ	*--vc-max-reads-per-raw-region*	200000
ミトコンドリア	*--vc-target-coverage-mito*	40000
ミトコンドリア	*--vc-max-reads-per-active-region-mito*	40000
ミトコンドリア	*--vc-max-reads-per-raw-region-mito*	40000

以下の例は、バリアント記録でレポートされるDPが、生殖細胞系列モードで*--vc-target-coverage*の初期設定値500を超える場合があることを示しています：

例えば、*--vc-target-coverage*の初期設定値が500であると仮定します。位置1で開始するリードが400、位置2で開始する別のリードが400、および位置3で開始する別のリードが400存在する場合、ターゲットカバレッジオプションは適用されません（400 < 500であるため）。位置4にバリアントが存在する場合、バリアントのレポートされる深度は最大で1200になる可能性があります。この例は、バリアント記録でレポートされるDPが、*--vc-target-coverage*の値を超える場合があることを示しています。

ターゲットカバレッジステップが終了すると、同じ位置を共有するリードの最大数が500になります（*--vc-target-coverage*が500に設定されている場合）。

次のダウンサンプリングステップは、*--vc-max-reads-per-raw-region*および*--vc-max-reads-per-active-region*の制限を適用することです。このステップでは、同じ位置を共有するリードの最大数を、最初のステップでの最大値500からさらに減らすことができます。これらのオプションは、水平化ダウンサンプリングメソッドを使用して、全領域のリードの合計数を制限するのに使用します。

ダウンサンプリングメカニズムにより領域の開始境界から各開始位置をスキャンし、その位置から1つのリードを破棄してから次の位置に移動し、この操作をリードの合計数が閾値未満になるまで行います。全領域でリードの合計数を閾値未満にするには、全領域でいくつかの工程が必要になる可能性があります。閾値の条件が満たされると、領域で最後に処理された位置に関係なく、ダウンサンプリングステップが停止します。

開始位置が同じである任意の位置でのリード数が*--vc-min-reads-per-start-pos*以下である場合、その位置をスキップして、任意の開始位置でのリード数が必ず最小数（*--vc-min-reads-per-start-pos*に設定）以上であることを確認します。

ダウンサンプリングが発生すると、保持または除去するリードの選択がランダムになります。ただし、乱数ジェネレーターのシードを初期設定値に設定して、ランごとにジェネレーターが同じセットの値を生成していることを保証します。これにより再現性の高い結果が保証されます。これは同じインプットデータを使用した場合はラン間で変動が生じないということを意味しています。

gVCF出力

ゲノムVCF（gVCF）ファイルには、バリアントに関する情報とリファレンスゲノムへのホモ接合性を維持していると決定された位置に関する情報が含まれています。ホモ接合性領域では、gVCFファイルには、バリアントまたは別のアリルがないことをリードがどれだけよく裏付けているのかを示す統計値が含まれています。gVCFファイルには、人為的な<NON_REF>アリルが含まれています。リファレンスおよびバリアントに対応していないリードには、<NON_REF>アリルが割り当てられます。DRAGENではこれらのリードを使用して、コールされないままにするのではなく、ホモ接合性リファレンスとして位置をコールできるのかどうかを判断します。結果として生じるスコアは、ホモ接合性リファレンスコールにおけるPhred値レベルの信頼度を表しています。生殖細胞系列モードでのスコアはFORMAT/GQとなり、体細胞モードでのスコアはFORMAT/SQとなります。

以下は、使用できるgVCF出力オプションです。

オプション	説明
--vc-emit-ref-confidence	gVCF出力を有効にするには、GVCFに設定します。初期設定では、スコアが類似したホモ接合性リファレンスコールの連続したランは、ブロック（hom-refブロック）に分類されます。hom-refブロックは、ディスク空き容量および下流の解析ツールの処理時間を節約します。DRAGENでは、初期設定モードを使用することを推奨しています。 バンドなし出力を生成するには、--vc-emit-ref-confidenceをBP_RESOLUTIONに設定します。
--vc-enable-vcf-output	gVCFラン時にVCFファイル出力を有効にするには、trueに設定します。初期設定値はfalseです。
--vc-gvcf-bands	初期設定の--vc-emit-ref-confidence gvcf（バンド付きモード）を使用する場合、DRAGENは類似したGQまたはSQスコアで、gVCFの記録を分類します。初期設定では、カットオフ値は生殖細胞系列の場合1 10 20 30 40 60 80で、体細胞の場合1 3 10 20 50 80です。例えば、バンド[0, 10)、[10, 50)、および≥ 50を定義するには、--vc-gvcf-bands 10 50を使用します。

gVCFのすべてのエントリーが連続しているとは限りません。ファイルには、バリアントラインおよびhom-refブロックによりカバーされていないギャップが含まれている場合があります。ギャップは、コール可能ではない領域に対応しています。MAPQスコア0以上のリードが１つもその領域にマップされていない場合、領域はコール可能ではありません。

生殖細胞系列モードでは、コールの閾値はVCFの場合よりgVCFの場合の方が低くなります。gVCF出力は、同じサンプルにおいてVCFランとは異なる数のバリアントを示す場合があります。二対立遺伝子および複対立遺伝子のコール数が異なる可能性があります。これは、gVCFモードには可能性が高いわずか2つのアリルではなく、座位で可能性のあるすべてのアリルが含まれているためです。これは、VCFでの二対立遺伝子コールを、gVCFでの複対立遺伝子コールとして出力できることを意味しています。gVCFでの遺伝型は依然として可能性が高い2つのアリルを示しているため、バリアントコールは同じままです。

以下は、hom-refブロックコールとバリアントコールを含むgVCF記録の例です。

```
1 39224 . C <NON_REF> . PASS END=39260
GT:AD:DP:GQ:MIN_DP:PL:SPL:ICNT
0/0:2,0:2:3:1:0,3,37:0,3,37:3,0
1 39261 . T C,<NON_REF> 15.59 PASS
DP=3;MQ=12.73;MQRankSum=0.736;ReadPosRankSum=0.736;FractionInformativeRea
ds=1.000
GT:AD:AF:DP:F1R2:F2R1:GQ:PL:SPL:ICNT:GP:PRI:SB:MB
0/1:1,2,0:0.667,0.000:3:1,0,0:0,2,0:5:49,0,1,69,7,75:66,0,8:1,0:1.5592e+0
1,1.5915e+
00,5.5412e+00,7.0100e+01,4.3330e+01,8.0068e+01:0.00,34.77,37.77,34.77,69.
54,37.77:0,1,0,2:0,1,2,0
```

QUAL、QD、およびGQ

単一サンプルのVCFとgVCFでは、QUALはVCF仕様の定義に従います。VCF仕様の詳細については、samtools/hts-specs GitHubリポジトリーで入手可能な最新のVCF資料を参照してください。

- QUALは部位にバリアントがないPhred値の確率であり、以下のように計算します：

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

  つまり、QUAL = GP (GT=0/0) です。ここで、GP は Phred 値の事後遺伝型確率です。

  QUAL = 20 は、部位にバリアントが存在する確度が 99% であることを意味しています。また VCF ファイルでは、GP 値は Phred 値でも示されています。

- GQは、コールが誤っているPhred値の確率です。
  GQ = -10*log10(p) です。ここで、p はコールが誤っている確度です。

  GQ = -10*log10(sum(10.^(-GP(i)/10))) です。ここで、sum は得られなかった GT を引き継いでいます。

  GQ が 3 であるということはコールが誤っている可能性が 50% であることを示しており、GQ が 20 であるということはコールが誤っている可能性が 1% であることを示しています。

- QDは、リード深度DPでノーマライズしたQUALです。

以下の表に公式をまとめて示します。

メトリクス	QUAL	GQ	QD
説明	部位にバリアントがない確度	コールが誤っている確度	深度でノーマライズしたQUAL
公式	QUAL = GP(GT=0/0)	GQ = -10*log10(p)	QUAL/DP
スケール	符号なしPhred	符号なしPhred	符号なしPhred
数値例	QUAL = 20:部位にバリアントが存在しない可能性が1%である QUAL = 50:部位にバリアントが存在しない可能性が10万分の1である	GQ = 3、コールが誤っている可能性が50%である GQ = 20、コールが誤っている可能性が1%である	

フェージングおよびフェージングされたバリアント

DRAGENは、生殖細胞系列のVCFおよびgVCFファイルにおいて、フェージングされたバリアント記録の出力に対応しています。2つ以上のバリアントがともにフェージングされた場合、フェージング情報はサンプルレベルのアノテーションであるFORMAT/PSにコード化されます。FORMAT/PSは、フェージングされたバリアントが存在するセットを同定します。フィールドの整数値は、セット内で最初にフェージングされたバリアントの位置を表しています。PS値が一致する同じコンティグ内の記録はすべて同じセットに属しています。

```
##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">
```

以下は、2つのSNPがともにフェージングされている、DRAGEN単一サンプルgVCFの例です。

```
chr1 1947645 . C T,<NON_REF> 48.44 PASS
DP=35;MQ=250.00;MQRankSum=4.983;ReadPosRankSum=3.217;FractionInformativeR
eads=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+01,8.387e-
05,5.300e+01,4.500e+02,4.500e+02,4.500e+02:0.00,34.77,37.77,34.77,69.54,3
7.77:11,9,10,5:12,8,8,7:1947645

chr1 1947648 . G A,<NON_REF> 50.00 PASS
DP=36;MQ=250.00;MQRankSum=5.078;ReadPosRankSum=2.563;FractionInformativeR
eads=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+01,7.067e-
05,5.204e+01,4.500e+02,4.500e+02,4.500e+02:0.00,34.77,37.77,34.77,69.54,3
7.77:10,6,11,9:8,8,12,8:1947645
```

ジェノタイピングステップ時には、活性領域にわたってすべてのハプロタイプおよびすべてのバリアントが考慮されています。バリアントの各ペアについて、同じハプロタイプのすべてで両方のバリアントが存在するか、またはいずれかがホモ接合性バリアントの場合、これらのバリアントはともにフェージングされます。バリアントが異なるハプロタイプのみに存在する場合、これらのバリアントは互いに反対にフェージングされます。同じハプロタイプの一部（ただしその他ではない）にヘテロ接合性バリアントが存在する場合、フェージングは中止され、活性領域のフェージング情報は出力されません。

フェージングされたバリアントの組み合わせ

同じフェージングセットに属するフェージングされたバリアント記録は、単一のVCF記録に組み合わせることができます。例えば、位置chr2 115035のリファレンスがAであると仮定して、以下の2つのフェージングされたバリアントを組み合わせます。

```
chr2 115034 .G C GT:PS 0|1:115034
chr2 115036 .C T GT:PS 0|1:115034
```

フェージングされたバリアントを、以下のように組み合わせます。

```
chr2 115034 .GAC CAT GT:PS 0|1:115034
```

`--vc-combine-phased-variants-distance`コマンドラインオプションを使用して、組み合わされるフェージングされたバリアント間の最大距離を指定できます。初期設定値は0であり、オプションを無効にします。オプションを有効にすると、フェージングセットで指定した距離値内にあるフェージングされたバリアントをすべて組み合わせます。

Ploidyサポート

現在、スモールバリアントコーラーは、ミトコンドリアコンティグを除いて、リファレンス内のすべてのコンティグでploidy 1または2のみをサポートしており、ここでは連続したアリル頻度アプローチを使用しています（100 ページの 「ミトコンドリアコール」 を参照）。他のすべてのコンティグに対するploidy 1または2の選択は、以下のようにして決定します。

- コマンドラインで*--sample-sex*が指定されていない場合、Ploidy Estimatorが性を判別します。Ploidy Estimatorが性核型を判別できないか、または性染色体数的異常を検出した場合、すべてのコンティグはploidy 2で処理されます。
- コマンドラインで*--sample-sex*が指定されている場合、コンティグは以下のように処理されます。
  - Femaleサンプルの場合、DRAGENはすべてのコンティグをploidy 2で処理し、フィルターPloidyConflictを使用してchrYのバリアントコールをマーキングします。
  - Maleサンプルの場合、DRAGENは性染色体以外に対して、すべてのコンティグをploidy 2で処理します。DRAGENはchrXをploidy 1で処理しますが、PAR領域は例外でploidy 2で処理されます。chrYは、全体にわたってploidy 1で処理されます。

DRAGENは、命名規則X/YまたはchrX/chrYにより性染色体を検出します。他の命名規則はサポートされていません。

--sample-sex	Ploidyの推定	スモールバリアントコーラーのサンプルの性
male	該当なし	Male
female	該当なし	Female
None	該当なし	None
auto(初期設定)	XY	Male
auto(初期設定)	XX	Female
auto(初期設定)	その他すべて	None

スモールバリアントコールでの重複メイト

DRAGENは、重複メイトを所定のイベントの独立した証拠として取り扱うのではなく、生殖細胞系列および体細胞パイプラインの両方において、以下のようにして処理します。

- 2つの重複メイトがアリルに関して最高のHMMスコアでお互いに一致している場合、ジェノタイパーは最大および2番目に高いHMMスコア間の最大差によりメイトを使用します。他のメイトのHMMスコアはゼロになります。
- 2つの重複メイトが一致していない場合、ジェノタイパーは2つのメイトのHMMスコアを合計し、組み合わせた結果に一致するメイトに組み合わせたスコアを割り当て、もう一方のメイトのHMMスコアをゼロに変更します。
- 重複メイトの塩基クオリティはこれ以上調整されません。

ミトコンドリアコール

通常、各哺乳類細胞には約100個のミトコンドリアが存在します。各ミトコンドリアは、2～10コピーのミトコンドリアDNA（mtDNA）を持ちます。例えば、chrMコピーの20%がバリアントを有している場合、アリル頻度（AF）は20%です。これは、連続したアリル頻度とも呼ばれます。chrMでのバリアントのAFが、0%～100%の間のいずれかであると予測されます。

DRAGENは、連続したAFパイプラインを通してchrMを処理します。この場合、単一のALTアリルについて考慮し、AFを推定します。推定したAFは、0%～100%の間のいずれかです。

QUALは、chrMバリアント記録の出力ではありません。代わりに、信頼度スコアがFORMAT/SQになり、所定の座位にバリアントが存在するPhred値の信頼度を提供します。

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

GQは、chrMバリアント記録の出力ではありません。これは、DRAGENが複数の二倍体遺伝型の候補を検出しないためです。代わりに、ALTアリルが候補のアリルであるとみなされます。FORMAT/SQ > `vc-sq-call- threshold`（初期設定 = 3.0）の場合、FORMAT/GTは`0/1`に設定されます。FORMAT/AFはバリアントアリル頻度の推定値を生成しますが、これは[0,1]範囲のいずれかです。アリル頻度が非常に高い場合（AF ≥ 95%）、FORMAT/GTは`1/1`に設定されます。

ミトコンドリアバリアントコールに以下のフィルターを適用できます。

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

VCF のコール出力をする場合は、SQ 閾値を設定します。初期設定は 0.1 です。

- *--vc-sq-filter-threshold*
  出力した VCF コールをフィルタリング済みとしてマーキングする場合は、SQ 閾値を設定します。初期設定は 3.0 です。

- *--vc-enable-triallelic-filter*
  複対立遺伝子フィルターを有効にします。初期設定値は false です。

- FORMAT/SQ < `vc-sq-call-threshold`の場合、バリアントはVCFの出力ではありません。
- FORMAT/SQ > `vc-sq-call-threshold`であるがFORMAT/SQ < `vc-sq-filter-threshold`の場合、バリアントはVCFに出力されますが、FILTER=weak_evidenceです。
- FORMAT/SQ > `vc-sq-call-threshold`、FORMAT/SQ > `vc-sq-filter-threshold`、および他のフィルターが設定されていない場合、バリアントはVCFに出力され、FILTER = PASSです。

以下は、chrMのVCF記録の例です。これは、1つのコールはAFが非常に高く、もう1つのコールはAFが非常に低い例です。両方の場合において、FORMAT/SQ > `vc-sq-call-threshold`です。FORMAT/SQ > `vc-sq-filter-threshold`でもあるため、FILTERアノテーションはPASSです。

```
chrM 513 . GCA G . PASS DP=4937;MQ=235.28;FractionInformativeReads=0.883
GT:SQ:AD:AF:F1R2:F2R1:DP:SB:MB
1/1:95.46:33,4327:0.992:7,1081:26,3246:4360:31,2,2371,1956:10,23,2811,151
6
chrM 1713 . A G . PASS DP=7175;MQ=165.91;FractionInformativeReads=0.995
GT:SQ:AD:AF:F1R2:F2R1:DP:SB:MB
0/1:21.49:7122,20:0.003:3066,10:4056,10:7142:3896,3226,8,12:3605,3517,10,
10
```

FORMAT/GT

NON_REF領域では、FORMAT/GTは0/0にハードコードされており、バリアント座位では、FORMAT/GTは0/1にハードコードされています。

chrMのFORMAT/GTは、FORMAT/AFでは決定されません。FORMAT/GTは、バリアントがある位置で放出されているかどうかで決定されます。バリアントを放出するかどうかは、FORMAT/SQスコアを閾値と比較することにより決定されます。FORMAT/SQ > `vc-sq-call-threshold`のバリアントはすべて出力されます。FORMAT/SQ < `vc-sq-call-threshold`のバリアントはgVCFでは放出されず、FORMAT/GT = 0/0のNON_REF領域にバンド付けされます。

所定の位置では、以下の2つのシナリオのうちの1つが発生します。

- バリアントは所定の位置でジェノタイパーにより検出され、FORMAT/SQ > `vc-sq-call-threshold`となります。この場合、FORMAT/GTは0/1にハードコードされます。DRAGENは、その位置で計算したFORMAT/AD、FORMAT/DP、およびFORMAT/AFを出力します。
- バリアントが所定の位置では検出されないか、またはバリアントは検出されますが、FORMAT/SQ < `vc-sq- call-threshold`となります。この場合、FORMAT/GTは0/0にハードコードされます。位置が同じシナリオ内にある場合、位置は連続した位置によりバンド付けされます。FORMAT/GT = 0/0であるすべての位置はともにバンド付けされます。バンドでレポートされるFORMAT/DPは、バンド内のすべての位置にわたる中央値DPとして計算されます。バンドでレポートされるFORMAT/AD値は、FORMAT/DP = 中央値DPの位置で選択されたAD値です。ジョイントVCFでは、FORMAT/AFはFORMAT/ADに基づいて計算されます。

以下は、トリオジョイントVCFのchrMにおけるバリアント記録の例です。バリアントは、フィルター閾値を合格した信頼度スコアにより、2番目のサンプルで検出されました。最初および3番目のサンプルではGT = 0/0であり、これは暫定的なhom-refコールを示しています（つまり、サンプルの位置が、十分な信頼度でバリアントが検出されなかったNON_REF領域内にある）が、weak_evidenceフィルタータグはこのコールが低い信頼度で検出されていることを示しています。

```
chrM 2623 . A G . . DP=18772;MQ=111.77 GT:AD:AF:DP:FT:SQ:F1R2:F2R1
0/0:6841,7:0.001:4334:weak_evidence:0:.:.
0/1:6736,2053:0.234:8789:PASS:21.32:3394,1060:3342,993
0/0:6086,9:0.001:5613:weak_evidence:0:.:.
```

重複バリアントのジョイント検出

単一の活性領域の複数の座位にあるバリアントが一緒に検出された場合、ジェノタイピングが有効なことがあります。以下の条件が満たされている場合、DRAGENは座位を1つのジョイント検出領域に組み合わせます：

- 座位に、お互いに重複するアリルがある。
- 座位がSTR領域にあるか、またはSTR領域から10塩基未満の間隔である。
- 座位が、お互いに10塩基未満の間隔である。

ジョイント検出では、ジョイント検出領域のアリルの可能性のあるすべての組み合わせが表されているハプロタイプリストを生成します。この計算により、ハプロタイプの数が大きくなります。ジェノタイピング時、ジョイント検出では、観察されたリードが集積している前提で、各ハプロタイプペアが真である尤度を計算します。遺伝型の尤度は、遺伝型でアリルをサポートしているハプロタイプペアの尤度の合計として計算されます。最大尤度の遺伝型がレポートされます。

ジョイント検出を有効にするには、`--vc-enable-joint-detection`をtrueに設定します。ジョイント検出を有効にすると、ランタイムが多少増加します。

ROHコーラー

ホモ接合性領域（ROH）は、スモールバリアントコーラーの一部として検出されます。コーラーは、常染色体ヒト染色体の全ゲノムコールからホモ接合性のランを検出して出力します。コマンドラインで指定するか、またはPloidy Estimatorの決定に従い、サンプル性核型がXXでない限り、性染色体は無視されます。下流のツールはROH出力を使用して、発端者と両親との間の血縁関係をスクリーニングして予測します。

領域は、これらのバリアント間に大きなギャップのない染色体の連続したバリアントコールとして定義されます。言い換えれば、領域は染色体かまたはSNVコールのない大きなギャップにより破棄されます。ギャップサイズは、3M塩基に設定されます。

ROHアルゴリズム

ROHアルゴリズムは、スモールバリアントコールで実行されます。アルゴリズムでは、複対立遺伝子部位があるバリアント、Indel、複雑なバリアント、PASSしていないフィルタリングコール、およびホモ接合性リファレンス部位は除外されています。ブロックリストBEDを使用してバリアントコールをさらにフィルタリングし、最終的にブロックリストフィルタリング後に深度フィルタリングが適用されます。フィルタリングしたコールの画分の初期設定値は0.2ですが、これはDP値で最高10%および最低10%のコールをフィルタリングします。アルゴリズムは、生成されたコールを使用して領域を検出します。

ROHアルゴリズムでは最初に、ヘテロ接合性SNVのない少なくとも50の連続したホモ接合性SNVコール、またはバリアント間の500,000塩基のギャップを含むシード領域を検出します。領域は、以下のように機能するスコアリングシステムを使用して拡張できます。

- スコアが、すべての追加ホモ接合性バリアントで増大し（0.025）、すべてのヘテロ接合性SNVの大きいペナルティで減少する（1～0.025）。これにより、領域のヘテロ接合性SNVの存在の許容度を提供します。
- 各領域は、領域が染色体の端に到達するか、SNV間で500,000塩基のギャップが存在するか、またはスコアが非常に低くなる（0）まで両端で拡張されます。

重複領域は、単一の領域に結合されます。単一の領域が、最初の領域の最初から2番目の領域の最後までギャップなしにコールされている場合、SNV間の500,000塩基のギャップにわたって領域を結合できます。領域に最大サイズは存在しませんが、領域は常に染色体境界で終了します。

ROHオプション

- *--vc-enable-roh*：trueに設定してROHコーラーを有効にします。ROHコーラーは、ヒト常染色体でのみ初期設定で有効です。falseに設定して無効にします。
- *--vc-roh-blacklist-bed*：指定されている場合、ROHコーラーは除外BEDファイルの領域に含まれているバリアントを無視します。DRAGENは、定評のあるすべてのヒトゲノムの除外BEDファイルを分配し、使用しているゲノムに一致するファイルを自動的に選択します。このオプションを明示的に使用しない場合は、ファイルを選択します。

ROH出力

ROHコーラーは、各行がホモ接合性の1つの領域を表している`<output-file-prefix>.roh.bed`という名前のROH出力ファイルを生成します。BEDファイルには以下の列が含まれています。

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

- スコアは、ホモ接合性およびヘテロ接合性バリアントの数の関数であり、ここでは各ホモ接合性バリアントはスコアが0.025ずつ増大し、各ヘテロ接合性バリアントではスコアが0.975ずつ減少します。
- 開始および終了位置は、0ベースの半開区間です。
- #Homozygousは、領域内のホモ接合性バリアントの数です。

- #Heterozygousは、領域内のヘテロ接合性バリアントの数です。

コーラーは、`<output-file-prefix>.roh_metrics.csv`という名前のメトリクスファイルも生成します。このファイルには、大きいROHの数および大きいROH（> 3 MB）内のSNPの割合がリストされています。

bアリル頻度出力

生殖細胞系列および体細胞VCFおよびgVCFランでは、初期設定でbアリル頻度（BAF）出力が有効です。

BAF値は、AFまたは(1 – AF)に等しくなります。ここで、

- AF = (alt_count / (ref_count + alt_count))
- BAF = 1 – AF、ただし、ref塩基 < alt塩基、塩基の優先度の順序がA < T <G < C < Nの場合のみ

厳密に1つのSNPオルタナティブアリルを持つ各スモールバリアントVCFエントリーの場合、出力にはBAF出力ファイル内に対応するエントリーが含まれています。

- <NON_REF>行は除外されます。
  - バリアントのINFOフィールドに「NML」タグも含まれていない場合、ForceGTバリアント（INFOフィールドで「FGT」タグによりマーキング）は出力に含まれていません。
  - ref_countおよびalt_countが両方ゼロであるバリアントは、出力に含まれていません。

BAFオプション

--vc-enable-baf
: b アリル頻度出力を有効または無効にします。初期設定は有効です。

BAF出力

BFが生成するのは、`<output-file-prefix>.baf.bw`および`<output-file-prefix>.hard-filtered.baf.bw`という名前のBigWig圧縮ファイルです。ハードフィルタリングファイルには、VCFで定義したフィルターを合格したバリアントのエントリーのみが含まれています（つまり、PASSエントリー）。

各エントリーには以下の情報が含まれています：

```
Chromosome Start End BAF
```

ここで：

- Chromosomeは、リファレンスコンティグに一致する文字列です。
- StartおよびEnd値は、0ベースの半開区間です。
- BAFは、浮動小数点値です。

体細胞モード

DRAGEN体細胞パイプラインでは、次世代シーケンサー（NGS）データの超高速解析により、体細胞染色体内のがん関連の変異を同定します。DRAGENでは、体細胞バリアント、生殖細胞系列バリアント、およびさまざまな系統的ノイズアーティファクトの可能性を考慮した確度モデルを使用して、一致したTumor-NormalペアおよびTumor-onlyサンプルの両方からSNVおよびIndelをコールします。体細胞バリアントを考慮する場合、DRAGENはploidy仮定を生成しないため、低頻度アリルの検出が可能になります。腫瘍サンプルのカバレッジが最大100xの座位の場合、DRAGENの検出の限界は5%のバリアントアリル頻度です。限界は座位ごとの深度の増加に合わせて調整されており、カバレッジが100xを超えて倍増するたびに半分になります。

Tumor-Normalパイプラインの場合、両方のサンプルが一緒に解析されます。DRAGENでは、生殖細胞系列バリアントおよび系統的なノイズアーティファクトが両方のサンプルで共有されている一方、体細胞バリアントは腫瘍サンプルにのみ存在すると仮定しています。体細胞バリアントのみがレポートされます。系統的なノイズアーティファクトを検出するために、DRAGENでは、正常サンプルでのカバレッジを腫瘍サンプルでのカバレッジの少なくとも半分にすることを推奨しています。

複数のフィルタリングステップの後、出力がVCFファイルとして生成されます。フィルタリングステップに不合格のバリアントは、出力VCFに保持されています。バリアントには、不合格だったフィルタリングステップを示すFILTERアノテーションが含まれています。

体細胞クオリティ（SQ）を主要なメトリクスとして使用して、コーラーが体細胞コールを生成した信頼度を記述できます。SQは、腫瘍サンプルのフォーマットフィールドとしてレポートされたPhred値の事後確率です。SQスコアがSQフィルター閾値未満のバリアントは、`weak_evidence`タグを使用してフィルタリングで除外されます。特異性に対して感度をトレードオフするには、SQフィルター閾値を調整します。閾値が低いと高感度のコーラーが生成され、閾値が高いと保存的なコーラーが生成されます。Tumor-Normal解析を実行する場合、正常サンプルのSQフィールドには、推定コールが生殖細胞系列バリアントであるPhred値の事後確率が含まれています。

体細胞モードオプション

体細胞モードには、以下のコマンドラインオプションがあります：

オプション	説明

--tumor-fastq1
--tumor-fastq2
	FASTQファイルのペアを、マッパーアライナーおよび体細胞バリアントコーラーに入力します。これらのオプションとその他のFASTQオプションを使用して、Tumor-Normalモードで実行できます。例えば:
	`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`

オプション	説明

--tumor-fastq-list	FASTQファイルのリストを、マッパーアライナーおよび体細胞バリアントコーラーに入力します。これらのオプションとその他のFASTQオプションを使用して、Tumor-Normalモードで実行できます。例えば:
	`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 *--tumor-cram-input*	マッピングされたBAMまたはCRAMファイルを、体細胞バリアントコーラーに入力します。これらのオプションとその他のBAM/CRAMオプションを使用して、Tumor-Normalモードで実行できます。
--vc-min-tumor-read-qual	バリアントコールで考慮するリードの最小クオリティ(MAPQ)を指定します。初期設定値はTumor-Normal解析では3で、Tumor-only解析では20です。
--vc-callability-tumor-thresh	腫瘍サンプルのcallability閾値を指定します。体細胞コール可能領域レポートには、腫瘍カバレッジが腫瘍閾値を超えるすべての領域が含まれています。初期設定値は15です。体細胞コール可能領域レポートの詳細については、241 ページの「体細胞コール可能領域レポート」を参照してください。
--vc-callability-normal-thresh	存在する場合、正常サンプルのcallability閾値を指定します。該当する場合、体細胞コール可能領域レポートには、正常カバレッジが正常閾値を超えるすべての領域が含まれています。初期設定値は5です。体細胞コール可能領域レポートの詳細については、241 ページの「体細胞コール可能領域レポート」を参照してください。

オプション	説明
--vc-somatic-hotspots *--vc-use-somatic-hotspots*	`--vc-somatic-hotspots`を使用して、ホットスポット入力VCFを指定します。ホットスポットVCFのバリアントは、多数の体細胞変異が予測される位置を示しています。例えば、ホットスポット入力VCFは、COSMICから得ることができます。DRAGENジェノタイピングプライアは、VCFで指定したすべての位置で増大させられるため、対応するリードが少ないこれらの部位のそれぞれでバリアントをコールできます。`vc-somatic-hotspots`の値を指定していない場合、DRAGENは`/opt/edico/config/somatic_hotspots*`からリファレンス固有のホットスポットVCFファイルを自動的に選択します。`vc-somatic-hotspots` VCFを指定している場合、VCFは常に初期設定のホットスポットVCFを優先します。 ホットスポット機能を無効にするには、`vc-use-somatic-hotspots=false`を使用します。初期設定のVCFも指定したVCFも考慮されません。
--vc-hotspot-log10-prior-boost	`--vc-use-somatic-hotspots`を使用する場合は、`vc-hotspotlog10-prior-boost`を使用して調整のサイズを制御します。初期設定値は40 Phredの増大に対応する4です（log10スケール）。

オプション	説明
--vc-enable-liquid-tumor-mode *--vc-tin-contam-tolerance*	Tumor-Normal解析では、DRAGENは血液腫瘍モードを実行することにより、Tumor-in-Normal（TiN）コンタミネーションを考慮します。血液腫瘍モードは、初期設定で無効にされています。血液腫瘍モードを有効にすると、DRAGENは指定した最大許容度レベルまでTiNコンタミネーションの存在する中でバリアントをコールできます。`--vc-enable-liquid-tumor-mode`では、初期設定の最大コンタミネーションTiN許容度0.15で血液腫瘍モードが有効になります。初期設定の最大コンタミネーションTiN許容度を使用する場合、体細胞バリアントは、腫瘍サンプル中の対応するアリルの最大15%のアリル頻度で正常サンプル中に観察されると予測されます。`vc-tin-contam-tolerance`により血液腫瘍モードを有効にすると、最大コンタミネーションTiN許容度を設定できます。血液腫瘍モードは、血液生検と同等ではありません。血液腫瘍モードの血液腫瘍は、白血病のような血液学的がんのことを言います。血液腫瘍の場合、血液中に腫瘍が存在するため、血液を正常コントロールとして使用できません。通常、皮膚または唾液が、正常サンプルとして使用されます。ただし、皮膚および唾液サンプルには依然として血液細胞が含まれている場合があるため、一致した正常対照サンプルには腫瘍サンプルの痕跡が含まれており、体細胞バリアントが正常サンプル中に低頻度で観察されます。コンタミネーションが考慮されない場合、真の体細胞バリアントを抑制することにより、感度に重大な影響を与える場合があります。通常、血液腫瘍モードでは中程度の深度（例えば、100xT/ 40xN）のWGSまたはWESライブラリを使用しており、これらの種類の深さで検出される最小のVAFは約5%です。通常、リキッドバイオプシーでは、ターゲット遺伝子パネル（例えば、500遺伝子）を非常に高い処理前深度（> 2000～5000x）で使用し、UMIインデックスを使用して0.5% LoDまで下がったVAFで感度を有効にします。
--vc-override-tumor-pcr-params-with-normal	腫瘍および正常サンプルで異なるシーケンスシステムまたは異なるライブラリー調製メソッドを使用する場合、DRAGENではこのオプションを`false`に設定することを推奨しています。Tumor-Normalモードでは、DRAGENは腫瘍および正常サンプルのPCRエラーパラメーターを個別に推定します。初期設定では、DRAGENは両方のサンプルの解析において腫瘍サンプルのパラメーターを無視し、正常サンプルのパラメーターを使用します。この初期設定により、体細胞バリアント率が高い場合に発生する可能性のある腫瘍サンプルのエラー率の過大評価を防止します。例えば、体細胞バリアント率が生殖細胞系列バリアント率に等しい生殖細胞系列混合データセットでは、エラー率が発生する場合があります。

Unique Molecular Identifersのサポート

異なるUMI使用症例のバリアントコールを最適化するためには、以下の2つのバッチモードを使用します。両方のオプションともに初期設定はfalseです。それらのいずれかをtrueに設定することにより、UMIバリアントコールを有効にします。

- *--vc-enable-umi-solid*：分類後のカバレッジ率が約200～300Xおよびターゲットアリル頻度が5%以上の固形がんにおいて、VC UMI固形モードは最適化されています。
- *--vc-enable-umi-liquid*：リキッドバイオプシーパイプラインは、血液腫瘍モードと同等ではありません。リキッドバイオプシーパイプラインは通常の血液サンプルから開始して、血液中を循環する、腫瘍由来のcell-free DNAから低VAFの体細胞バリアントを探します。この種の検査では、組織からではなく血漿からの腫瘍のプロファイリング（診断/バイオマーカー同定）を有効にしますが、この場合侵襲的な生検が必要になります。分類後のカバレッジ率が約2,000～2,500Xおよびターゲットアリル頻度が0.4%以上の血液生検パイプラインにおいて、VC UMI血液モードは最適化されています。

これらのバッチ設定には、vc-systematic-noiseフィルターは含まれていません。DRAGENでは、フィルターを追加することを推奨しています。詳細については、114 ページの 「系統的なノイズフィルタリング」 を参照してください。

UMIパイプラインでは、DRAGENは、サンプルごとにシーケンスシステムの上流に存在する可能性のある酸化および脱アミノ化アーティファクトのような、ヌクレオチド変異バイアスを推定できます。バリアントコール時に、DRAGENはバイアスを修正します。

この機能を有効にするには、`--vc-target-bed`を使用しますが、この場合バリアントコールのターゲット領域も指定します。つまり、`--vc-snp-error-cal-bed`です。`--vc-snp-error-cal-bed`を使用する場合、ターゲットBEDファイルと異なる場合またはターゲットBEDファイルが指定されていない場合、ヌクレオチド置換バイアスを推定する領域を指定します。サードパーティーツールを使用して分類リードを生成する場合、ベースコールクオリティスコアでシーケンスシステムからのエラーのみを定量するように、ツールを設定します。DRAGENではこのエラー推定を使用して、シーケンスシステムの上流のエラーを考慮します。

体細胞コーリング後フィルタリング

オプション

以下のオプションは、体細胞コーリング後フィルタリングで使用できます：

オプション	説明
--vc-sq-call-threshold	VCFでコールを出力します。初期設定はTumor-Normalでは3.0で、Tumor-onlyでは0.1です。*vc-sq-filter-threshold*の値が*vc-sq-call-threshold*より低い場合、コール閾値の代わりにフィルター閾値を使用します。
--vc-sq-filter-threshold	出力したVCFコールをフィルタリング済みとしてマーキングします。初期設定はTumor-Normalでは17.5で、Tumor-onlyでは3.0です。
--vc-enable-triallelic-filter	複対立遺伝子フィルターを有効にします。初期設定はtrueです。

オプション	説明
--vc-enable-af-filter	アリル頻度フィルターを有効にします。初期設定値はfalseです。trueに設定すると、VCFは、アリル頻度がAFコール閾値より低いバリアント、またはアリル頻度がAFフィルター閾値より低く、低AFフィルタータグでタグ付けされているバリアントを除外します。初期設定のAFコール閾値は1%で、初期設定のAFフィルター閾値は5%です。閾値を変更するには、コマンドラインオプション`vc-af-call-threshold`および`vc-af-filter-threshold`を使用します。
--vc-enable-non-homref-normal-filter	non-homref正常フィルターを有効にします。初期設定値はtrueです。trueに設定すると、正常サンプル遺伝型がホモ接合性リファレンスではない場合、VCFはバリアントをフィルタリングして除外します。
--vc-enable-vaf-ratio-filter	`alt_allele_in_normal`フィルターで除外される1つの条件を追加します。初期設定値はfalseです。trueに設定すると、正常サンプルAFが腫瘍サンプルAFの20%より大きい場合、VCFはバリアントをフィルタリングして除外します。

フィルター

以下の表は、使用できる体細胞コールフィルターを示しています。

体細胞モード	フィルターID	説明
Tumor-onlyとTumor-Normal	clustered_events	クラスターイベントが所定の活性領域で観測されました。クラスターイベントの閾値は設定可能です（初期設定≥ 3）。`--vcenable-gatk-acceleration=true`を使用している場合にのみ有効です。
Tumor-onlyとTumor-Normal	weak_evidence	バリアントが尤度の閾値を満たしていません。尤度比は、SQ Tumor-Normalでは< 17.5で、SQ Tumor-onlyでは< 3.0です。
Tumor-onlyとTumor-Normal	multiallelic	腫瘍のこの位置に2つ以上のALTアリルが存在する場合、フィルタリングされる部位。

体細胞モード	フィルターID	説明
Tumor-onlyとTumor-Normal	str_contraction	ALTアリルがリファレンスより繰り返しユニットが1だけ小さい場合に、疑われるPCRエラー。*--vc-enable-gatk-acceleration=true*を使用している場合にのみ有効です。
Tumor-onlyとTumor-Normal	base_quality	この座位でのALTリードの中央値塩基クオリティは< 20です。
Tumor-onlyとTumor-Normal	mapping_quality	この座位でのALTリードの中央値マッピングクオリティは< 20（Tumor-Normal）または< 30（Tumor-only）です。
Tumor-onlyとTumor-Normal	fragment_length	所定の座位でのaltリードの断片化長中央値とrefリードの断片化長中央値の絶対差が> 10000。
Tumor-onlyとTumor-Normal	read_position	リードの開始/終了および所定の座位間の距離の中央値< 5（バリアントがすべてのリードの端に接近しすぎている）。
Tumor-onlyとTumor-Normal	low_af	アリル頻度は、*--vc-af-filter-threshold*で指定した閾値未満です（初期設定は5%）。*--vc-enable-af-filter=true*を使用している場合にのみ有効です。
Tumor-onlyとTumor-Normal	systematic_noise	Tumor-NormalのAQスコアが< 10（初期設定）またはTumor-onlyのAQスコアが< 60（初期設定）の場合、部位はフィルタリングされます。
Tumor-onlyとTumor-Normal	low_frac_info_reads	有用なリードの断片が閾値未満です。閾値の初期設定値は0.5です。
Tumor-onlyとTumor-Normal	low_depth	リードの数が低すぎるため、部位がフィルタリングされました。初期設定ではフィルターはoffです。

Tumor-NormalとTumor-only

体細胞モード	フィルターID	説明
Tumor-Normal	noisy_normal	9.9%を超えるアリル頻度において、正常サンプル中で3つを超えるアリルが観察されています。
Tumor-Normal	alt_allele_in_normal	正常サンプルのALTアリル頻度は、0.2 +最大コンタミネーション耐性を超えています。固形がんモードでは、値は0です。血液腫瘍モードでは、初期設定値は0.15です。オプションの条件については、`vc-enable-vaf-ratio-filter`を参照してください。
Tumor-Normal	filtered_reads	リードの90%超がフィルタリングで除外されました。
Tumor-Normal	no_reliable_supporting_read	腫瘍サンプル中に、信頼できる対応リードが見つかりませんでした。信頼できる対応リードは、マッピングクオリティ≥ 40、断片化長≤ 10000、ベースコールクオリティ≥ 25、およびリードの開始/終了からの距離≥ 5であるaltアリルに対応するリードです。
Tumor-Normal	too_few_supporting_reads	バリアントには、腫瘍サンプル中の< 3リードで対応しています。
Tumor-Normal	non_homref_normal	正常サンプル遺伝型は、ホモ接合性リファレンスではありません。
Tumor-Normal	germline_risk	正常サンプル中にアリルが存在する尤度は> 0.025です。*--vc-enable-gatk-acceleration=true*を使用している場合にのみ有効です。
Tumor-Normal	artifact_in_normal	正常リードセットのTLOD（正常アーティファクトLOD）は> 0.0です。正常サンプル中のアリル断片が腫瘍中のアリル断片より大幅に小さい場合（normalAlleleFraction < (0.1 * tumorAlleleFraction)）に、コールされない正常アーティファクト。*--vc-enable-gatk-acceleration=true*を使用している場合にのみ有効です。

QUALは、体細胞バリアント記録の出力ではありません。代わりに、信頼度スコアはFORMAT/SQです。

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

フィールドはサンプルに固有です。腫瘍サンプルの場合、所定の座位に体細胞バリアントが存在する証拠を、フィールドにより定量します。

正常サンプルも使用できる場合、所定の座位で正常サンプルがホモ接合性リファレンスである証拠を、対応するFORMAT/SQ値により定量します。

GQは、体細胞バリアント記録の出力ではありません。これは、DRAGENが複数の二倍体遺伝型の候補を検査しないためです。代わりに、ALTアリルが体細胞バリアントの候補であるとみなされます。腫瘍SQ > `vc-sq-call-threshold`（初期設定は3）の場合、腫瘍サンプルのFORMAT/GTは0/1にハードコードされます。また、FORMAT/AFは体細胞バリアントアリル頻度の推定値を生成しますが、これは[0,1]の範囲内のいずれかです。

- 腫瘍SQ < `vc-sq-call-threshold`の場合、VCFでバリアントは出力されません。
- 腫瘍SQ > `vc-sq-call-threshold`であるが腫瘍SQ < `vc-sq-filter-threshold`の場合、バリアントはVCFで出力されますが、FILTER=weak_evidenceです。
- 腫瘍SQ > `vc-sq-call-threshold`で腫瘍SQ > `vc-sq-filter-threshold`の場合、バリアントはVCFで出力され、FILTER=PASSです（バリアントが別のフィルターでフィルタリングされていない場合）。
- 初期設定のvc-sq-filter-thresholdはTumor-Normal解析では17.5で、Tumor-only解析では3.0です。

以下は、体細胞T/N VCF記録の例です。腫瘍SQ > `vc-sq-call-threshold`であるが腫瘍SQ < `vc-sq-filter-threshold`の場合、FILTERはweak_evidenceとしてマーキングされます。

```
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
```

クラスターイベントのペナルティは、バリアント出力における上の規則に対する例外です。初期設定では、Tumor-Normalモードにおいて、クラスターイベントペナルティはクラスターイベントフィルターの代わりです。DRAGENは、ともにクラスター化されるイベントが多すぎる際にハードフィルターを適用するのではなく、正相のクラスターイベントのSQスコアにペナルティを適用します。証拠が乏しいクラスターイベントはこれ以上コールされませんが、証拠が有力なクラスターイベントは依然としてコールできます。これは、クラスター正相バリアントを観測する事前確率を下げることと同等です。バリアントを出力することを決定した後にペナルティを適用することにより、ペナルティを適用しないスコアが十分に高い場合、ペナルティを適用するバリアントがVCFに依然として表示されます。

gVCF出力

Tumor-onlyデータセットのgVCFファイルを出力できます。gVCFファイルには、入力ゲノムの位置ごとに1つの記録が含まれています。その位置でバリアントがコールされなかった場合、DRAGENは新たに<NON_REF>アリルを作成します。ホモ接合性リファレンス（homref）コールを裏付けないリードには、<NON_REF>アリルが割り当てられます。体細胞モードでは、任意に低いアリル頻度で存在している可能性があるバリアントは、検出不能です。体細胞homrefコールでは、位置で任意のアリル頻度の体細胞バリアントが存在しないことを保証できません。代わりに、検出レベル（LOD）以上で特定のアリル頻度を持つ体細胞バリアントが存在しない場合、DRAGENはその位置をホモ接合性リファレンスであるとみなします。LOD値が低い場合、少数のhomrefコールが生成されます。LOD値が高い場合、多数のhomrefコールが生成されます。

初期設定ではLODは5%に設定されていますが、`--vc-gvcf-homref-lod`オプションを使用して別の値を入力できます。

系統的なノイズフィルタリング

DRAGENを体細胞モードで使用すると、部位固有のノイズレベルによりBEDファイルを指定してシーケンス/系統的なノイズをフィルタリングして除外できます。部位固有のノイズレベルを使用して、Phred値と類似したAQスコアを計算します。AQスコアが定義した閾値より小さい場合、バリアントは系統的なノイズとしてフィルタリングされます。腫瘍サンプルを採取した被験者に必ずしも一致しているとは限らない正常サンプルでのラン時に、DRAGEN体細胞パイプラインで生成したVCFを使用して系統的なノイズBEDファイルを構築します。ファイルには、数ダースのサンプルが含まれている場合があります。理想的には、サンプルを同じライブラリー調製キットとシーケンスシステムで収集した正常サンプルにすることにより、ライブラリー調製またはシーケンシング時に発生する系統的なエラーが存在する場合、そのエラーを系統的なノイズBEDファイルに取り込みます。

以下は、使用できる系統的なノイズコマンドラインオプションです：

- *--vc-systematic-noise*：系統的なノイズBEDファイルを指定します。体細胞バリアントがAQ閾値に合格しない場合、バリアントは出力VCFのFILTER列に`systematic_noise`としてマーキングされます。
- *--vc-systematic-noise-filter-threshold*：AQ閾値を設定します。初期設定では、閾値はTumor-Normalでは10で、Tumor-onlyでは60です。

WGSおよびWESのいくつかの事前に構築された系統的なノイズファイルをダウンロードできます。最高性能を実現するには、同じライブラリー調製キットとシーケンスシステムで収集した正常サンプルを使用して、系統的なノイズBEDファイルを構築する必要があります。

系統的なノイズBEDファイルの生成

ライブラリー調製、シーケンスシステム、およびパネルを使用して収集した正常サンプルから、系統的なノイズBEDファイルを生成できます。パネルシーケンシングを使用する場合、50サンプルを推奨します。

BEDファイルを生成するには、以下のようにします。

1. 正常サンプルごとにVCF出力を生成するには、*--vc-detect-systematic-noise*を`true`に設定した正常サンプルで、DRAGEN体細胞Tumor-onlyを実行します。

   --vc-detect-systematic-noise のみを使用して、ノイズファイルを生成します。オプションは、腫瘍サンプルの解析を目的とはしていません。

2. VCFと以下のオプションを使用して、BEDファイルを構築します。

   BaseSpace Sequence Hub DRAGEN CNV Baseline Builder アプリを使用して、クラウドに系統的なノイズ BED ファイルを構築することもできます。

オプション	説明
--vc-systematic-noise-raw-input-list	入力VCFのリスト。行ごとに1つのVCFを入力します。

オプション	説明
--vc-systematic-noise-germline-vaf-threshold	系統的なノイズファイル構築から可能性のある生殖細胞系列を除去するための最小VAF。VAFが閾値よりも大きいバリアントは、系統的なノイズとはみなされません。初期設定は指定されておらず、これはすべてのバリアントが使用されていることを示します。 小さなパネルを使用する場合、推奨する閾値は0.3です。
--vc-systematic-noise-use-germline-tag	DRAGEN内部生殖細胞系列タグ付けを使用して、可能性のある生殖細胞系列を除去します。`--vc-systematic-noise-germline-vaf-threshold`と相互に排他的です。初期設定はfalseです。 WGSまたはWESを使用する場合、推奨する設定は`true`です。
--vc-systematic-noise-method	サンプルにわたって系統的なノイズレベル（ノイズアリル頻度）を計算するためのメソッド。`mean`を入力して平均ノイズアリル頻度を計算するか、`max`を入力して最大値を計算するか、または`aggregate`を入力してサンプルにわたる座位ごとの合計アリル/合計深度を計算します。初期設定は`mean`です。 WGSを使用する場合、推奨する設定は`max`です。WESを使用する場合、推奨する設定は`aggregate`です。高感度を実現するために小さなパネルを使用する場合、推奨する設定は`mean`または`aggregate`です。

事前に構築された系統的なノイズBEDファイル

WGSおよびWESの以下の事前に構築された系統的なノイズファイルは、DRAGEN Bio-It Platformサポートサイトページでダウンロードできます。

事前に構築された系統的なノイズ	コメント	正常サンプル数
WGS_hg38_v1.0_systematic_noise.bed.gz	WGS hg38	28サンプル。サンプルは、NovaSeq 6000でシーケンスされたIllumina DNA PCR-free kit、HiSeq Xでシーケンスされたillumina DNA PCR-free kit、およびHiSeq XでシーケンスされたTruSeq DNA Nano kitの混合物です。
WGS_hs37d5_v1.0_systematic_noise.bed.gz	WGS hs37d5	31サンプル。サンプルは、NovaSeq 6000でシーケンスされたIllumina DNA PCR-free kit、HiSeq Xでシーケンスされたillumina DNA PCR-free kit、およびHiSeq XでシーケンスされたTruSeq DNA Nano kitの混合物です。

事前に構築された系統的なノイズ	コメント	正常サンプル数
WGS_hg19_v1.0_systematic_noise.bed.gz	WGS hg19	31サンプル。サンプルは、NovaSeq 6000でシーケンスされたIllumina DNA PCR Free Kit、HiSeq XでシーケンスされたIllumina DNA PCR Free Kit、およびHiSeq XでシーケンスされたTruSeq DNA Nano kitの混合物です。
WES_Nextera_IDT_hg38_v1.0_systematic_noise.bed.gz	Nexteraライブラリー調製、IDT Exome、hg38	47サンプル。
WES_Nextera_IDT_hs37d5_v1.0_systematic_noise.bed.gz	Nexteraライブラリー調製、IDT Exome、hs37d5	47サンプル。
WES_Nextera_IDT_hg19_v1.0_systematic_noise.bed.gz	Nexteraライブラリー調製、IDT Exome、hg19	47サンプル。
WES_TruSeq_IDT_hg38_v1.0_systematic_noise.bed.gz	TruSeqライブラリー調製、IDT Exome、hg38	53サンプル。
WES_TruSeq_IDT_hs37d5_v1.0_systematic_noise.bed.gz	TruSeqライブラリー調製、IDT Exome、hs37d5	53サンプル。
WES_TruSeq_IDT_hg19_v1.0_systematic_noise.bed.gz	TruSeqライブラリー調製、IDT Exome、hg19	53サンプル。

複数サンプルのジョイント解析

DRAGENでは、複数サンプルのpedigreeベースおよび集団ベースのジョイント解析をサポートしています。pedigreeベースの解析では、同じ種由来のサンプルはお互いに関連付けられます。集団ベースの解析では、同じ種由来のサンプルはお互いに関連付けられません。

各サンプルのジョイント解析ではgVCFファイルが必要です。gVCFファイルを作成するには、`--vc-emit-ref-confidence gVCF`オプションで生殖細胞系列スモールバリアントコーラーを実行します。

gVCFファイルには、バリアントの位置に関する情報およびリファレンスゲノムへのホモ接合性と決定された位置に関する情報が含まれています。ホモ接合性領域では、gVCFファイルには、バリアントまたは別のアリルがないことにリードがどれだけ良好に裏付けているかを示す統計値が含まれています。信頼度が類似したレベルの塩基の連続したホモ接合性ランはブロックにグループ化され、hom-refブロックと呼ばれます。gVCFのすべてのエントリーが連続しているとは限りません。リファレンスには、バリアントラインおよびhom-refブロックによりカバーされていないギャップが含まれている場合があります。ギャップは、コール可能ではない領域に対応しています。MAPQスコア0より上のリードが少なくとも１つその領域にマップされていない場合、領域はコール可能ではありません。以下の例は、1つのサンプルがバリアントを有するジョイントVCF、および他の2つのサンプルがgVCFギャップに存在することを示しています。ギャップは「./.:.:.」で表されます。

```
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
```

pedigreeモード

pedigreeモードを使用して関連する個体からのサンプルを一緒に解析し、*de novo*コールを実行します。

pedigreeモードを起動するには、`--enable-joint-genotyping`オプションをtrueに設定します。パネル間の関係を記述するpedigreeファイルへのパスを指定するには、`--pedigree-file`オプションを使用します。

pedigreeファイルは、ファイル名が*.pedで終わるタブ区切りのテキストファイルにする必要があります。以下の情報が必須です。

列ヘッダー	説明
Family_ID	pedigree識別子。
Individual_ID	個体のID。
Paternal_ID	個体の父親のID。発端者の場合、値は0です。
Maternal_ID	個体の母親のID。発端者の場合、値は0です。
Sex	サンプルの性。Maleの場合、値は1です。Femaleの場合、値は2です。
Phenotype	サンプルの遺伝学的データ。不明の場合、値は0です。影響を受けない場合、値は1です。影響を受ける場合、値は2です。

以下は、入力pedigreeファイルの例です。

```
#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
```

De Novoコール

De Novo Callerはpedigree内のすべてのトリオを同定して、子ごとに*de novo*スコアを生成します。De Novo Callerは、単一のpedigree内の複数のトリオをサポートしています。pedigreeモードは、スモール、構造多型、およびコピー数バリアントの*de novo*コールをサポートしています。

pedigreeモードは、複数のステップで実行されます。以下は、FASTQ入力を使用しているトリオのワークフローの例です。

1. 単一サンプルアライメントおよびバリアントコールを実行し、pedigreeモードで以下の入力を使用してサンプルごとの出力を生成します。
   - スモールバリアントコーラーでのgVCFファイル。
   - コピー数コーラーでの*.tn.tsvファイル。
   - 構造多型コーラーでのBAMファイル。
2. スモールバリアントコーラーでpedigreeモードを実行します。
   詳細については、119 ページの「スモールバリアント De Novo コール」を参照してください。
3. コピー数コーラーでpedigreeモードを実行します。
   詳細については、161 ページの「マルチサンプル CNV コール」を参照してください。
4. 構造多型コーラーでpedigreeモードを実行します。
   詳細については、211 ページの「構造多型の de novo クオリティスコアリング」を参照してください。
5. De Novoバリアントスモールバリアントフィルタリングを実行します。
   詳細については、127 ページの「de novo スモールバリアントフィルタリング」を参照してください。

スモールバリアントDe Novoコール

Small Variant De Novo Callerは、一度にサンプルのトリオを考慮します。サンプルは、pedigreeファイルを介して関係しています。Small Variant De Novo Caller は、個別のサンプルgVCFからの遺伝型に基づいて、メンデル不一致が存在するすべての位置を決定します。Maleの性染色体は、二倍体として取り扱われているPAR領域以外一倍体として取り扱われます。

各位置はPedigreeコーラーを通して処理され、可能性のある遺伝型のジョイント後確率マトリクスが計算されます。確率を使用して、発端者がDQ信頼度スコアを持つ*de novo*バリアントを保有しているかどうかを判別します。3人の被験者はすべて、独立したエラーの起こり易さを持つと仮定されます。

gVCFからの元の遺伝型が2倍のメンデル不一致（例えば、0/0+0/0->1/1または1/1+1/1->0/0）を示す位置において、少なくとも1つのメンデル不一致が存在する最大のジョイント後確率に対して、トリオサンプルの遺伝型を調整できます。

DQの公式は、DQ = -10log10(1 - Pdenovo)です。

Pdenovoは、メンデル不一致の1つがあるジョイント後確率マトリクスのすべてのインデックスの合計です。

GT上書きステップでは、親のGTを上書きすることが可能です。複数のトリオの場合、親のGTは処理された最後のトリオに基づいています。トリオは、pedigreeファイルにリストされている順序で処理されます。現在DRAGENは、GTが上書きされた場合、VCFにアノテーションを追加しません。

マルチサンプルVCFファイルは、*de novo*クオリティスコアおよび関連する*de novo*コールを表すVCFファイルの出力に対して、FORMAT/DQおよびFORMAT/DNフィールドでアノテーションされます。VCFのDNフィールドは、各セグメントの*de novo*ステータスを示すために使用されます。

以下は、可能性のある値です：

- **Inherited**：コールされるトリオ遺伝型はメンデル遺伝と一致しています。
- **LowDQ**：コールされるトリオ遺伝型はメンデル遺伝と一致しておらず、DQは*de novo*クオリティ閾値未満です。
- **DeNovo**：コールされるトリオ遺伝型はメンデル遺伝と一致しておらず、DQは*de novo*クオリティ閾値以上です。

以下の例は、トリオのVCF行を示しています：

```
1 16355525. G A 34.46 PASS
AC=1;AF=0.167;AN=6;DP=45;FS=6.69;MQ=108.04;MQRankSum=0.156;QD=2.46;ReadPos
RankSum=0;SOR=0.01
GT:AD:AF:DP:GQ:FT:F1R2:F2R1:PL:GP:PP:DPL:DN:DQ
0/1:11,3:0.214:14:39:PASS:8,2:3,1:74,0,47:39.454,0.00053613,49.99:0,1,104:
74,0,47:DeNovo:0.6737
0/0:18,0:0:16:48:PASS:.:.:0,48,605:.:0,12,224:0,48,255:.:0/0:14,0:0:14:42:
PASS:.:.:0,42,490:.:0,5,223:0,42,255:.:.
```

De Novoスモールバリアントのオプション

以下のコマンドラインオプションは、*de novo*スモールバリアントコールで使用できます。

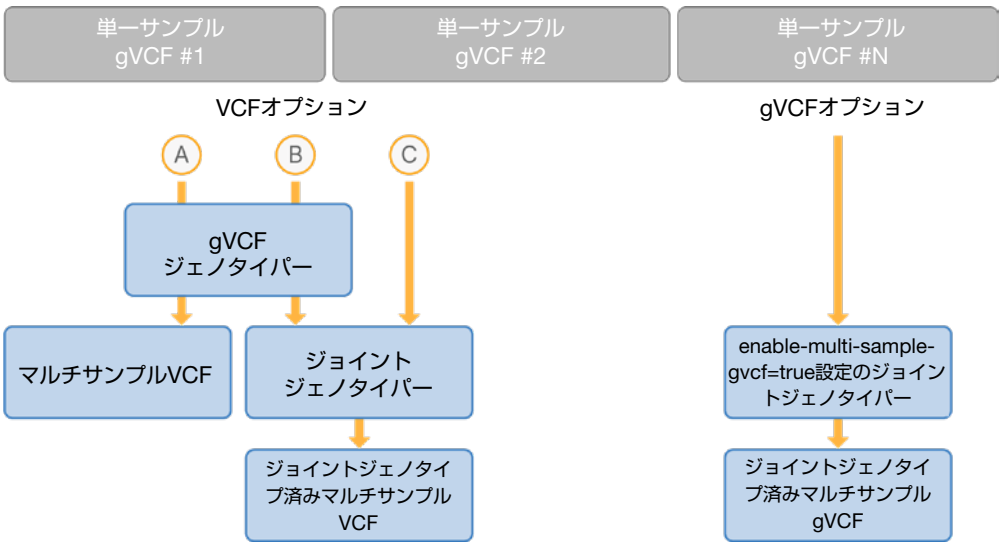
オプション	説明
--enable-joint-genotyping	ジョイントジェノタイピングコーラーを実行します。
--variant	ワークフローに対するgVCF入力を指定します。
--pedigree-file	トリオが存在する場合に、De Novoコールを有効にするpedigreeファイルを指定します。
--qc-snp-denovo-quality-threshold	*de novo*とみなされるSNPの最小DQ値を指定します。初期設定値は0.05です。
--qc-indel-denovo-quality-threshold	*de novo*とみなされるIndelの最小DQ値を指定します。初期設定値は0.02です。

集団モード

DRAGENは集団ベースの解析オプションを使用して、関連のない個体からのサンプルを一緒に解析します。集団モードを開始するには、以下のジェノタイパーを使用します。

- **gVCFジェノタイパー**：1セットの単一またはマルチサンプルgVCFを入力として使用し、入力gVCFのいずれかに見られるバリアントの1つのエントリーを含むマルチサンプルVCFを出力します。必要に応じて、hom-refブロックからの情報を使用し、すべての入力サンプルにわたってバリアントの遺伝型を判定します。gVCFジェノタイパーは、集団情報に基づいて遺伝型を調整しません。使用できるコマンドラインオプションの詳細については、122 ページの 「gVCFジェノタイパーのオプション」 を参照してください。
- **ジョイントジェノタイパー**：コホート全体からの情報を使用して、個別の遺伝型の精度を改善します。マルチサンプルVCF、マルチサンプルgVCF、または1セットの単一サンプルgVCFを入力できます。出力をマルチサンプルgVCFとして受け取るには、`--enable-multi-sample-gVCF`をtrueに設定します。使用できるコマンドラインオプションの詳細については、121 ページの 「ジョイントジェノタイパーのオプション」 を参照してください。

以下の図は、gVCFジェノタイパーとジョイントジェノタイパー間の異なるパスウェイおよびデータフローを示しています。



gVCFパスウェイは、サンプル数が3～15のpedigreeまたはコホートのような小さなデータセットにのみ適しています。VCFパスウェイは、大きいデータセットに調整できます。VCFパスウェイを使用する場合、単一のサーバーで約24時間で1,000サンプルを解析できます。

コホートに存在するバリアントおよび各コホートメンバーの該当するバリアントの遺伝型のリストを受け取るには、gVCFジェノタイパーを実行します。必要に応じて、2番目のマルチサンプルVCFの構築後にジョイントジェノタイパーを実行できます。ジョイントジェノタイパー出力は、集団情報に基づいてサンプルの遺伝型をより正確にします。gVCFジェノタイパー出力のみを使用して、バリアントに低深度または低遺伝型クオリティが含まれている場合、稀なバリアントをフィルタリングで除外してノイズを防止できます。出力ファイルでbcftoolsのようなオープンソースユーティリティを使用して、バリアントをフィルタリングします。

複数のpedigreeを比較するには、ジョイントジェノタイパーの出力でgVCFジェノタイパーを実行し、複数のジョイントコールpedigreeを単一のマルチサンプルVCFに結合します。

`--enable-multi-sample-gvcf=true` gVCFオプションを使用してジョイントジェノタイパーを設定し、マルチサンプルgVCFを記述します。

GATKからのgVCFファイルの処理

gVCFジェノタイパーおよびジョイントジェノタイパーは両方ともに、GATK v4.1を使用している際にGATKバリアントコーラーにより記述されたgVCFファイルを処理できます。このオプションを有効にするには、gVCFジェノタイパーおよびジョイントジェノタイパーの両方で`--vc-enable-gatk-acceleration=true`に設定します。

ジョイントジェノタイパーのオプション

このセクションでは、各ジェノタイパーで使用できるオプションに関する情報について説明します。

gVCFジェノタイパーのオプション

gVCFジェノタイパーは1セットの単一サンプルgVCFを使用して、入力gVCFのいずれかに見られるバリアントごとに1つのエントリーを含むマルチサンプルVCFを出力します。遺伝型は、集団情報では調整できません。

gVCFジェノタイパーは、S3バケットからのgVCFファイルも読み込めます。公開バケットのgVCFファイルでは、`--variant`または`--variant-list`においてプレフィックス*s3://*または*https://*の付いたURLを使用できます。バケットで認証が必要な場合、環境変数または設定ファイルを使用できます。htslib AWS S3プラグインの詳細については、Samtoolsウェブサイトを参照してください。

gVCFジェノタイパーは、gVCF入力ごとにインデックスファイルにアクセスする必要があります。各gVCFおよびインデックスファイルのURLは、`https://url1.gvcf.gz##idx##https://url2.gvcf.gz.tbi`として組み合わせてから、コマンドラインで*--variant*または*--variant-list*に渡す必要があります。

gVCFジェノタイパーでは、以下のパラメーターが使用できます。

オプション	説明
--enable-gvcf-genotyper	gVCFジェノタイパーを有効にするには、trueに設定します。
--ht-reference	FASTAフォーマットのリファレンスシーケンスが含まれているファイル。*--ht-reference*が必須です。
--output-directory	出力ディレクトリ。`--output-directory`が必須です。
--output-file-prefix	すべての出力ファイルにラベル付けするのに使用されるプレフィックス。`--output-file-prefix`が必須です。
--gg-output-format	出力ファイルフォーマット。初期設定値はvcf.gzです。許可されている出力ファイルフォーマットは、vcf.gz、vcf、またはbcfです。ジョイントジェノタイパーに対応しているのは、vcf.gzフォーマットのみです。別のフォーマットを使用する場合は、オープンソースのbcftoolsユーティリティを使用してフォーマットを変換する必要があります。
--gg-regions	gVCFジェノタイパーを実行する領域を指定するファイル。これらの領域外のバリアントは無視されます。ファイルは、bedファイルまたはchromosome:start-endを使用して指定したゲノム領域のリストにすることができます。ゲノム領域は、カンマまたは改行で区切る必要があります。エクソームまたは濃縮データを使用する場合、プローブで対象とした領域のリストを指定して、ターゲット領域の外にある信頼できない遺伝型バリアントを処理するのに要する追加の時間を制限します。
--gg-enable-concat	ゲノム領域の出力を単一の出力ファイルに連結します。初期設定では、値はtrueに設定されています。

オプション	説明
--gg-max-alternate-alleles	オルタナティブアリルの最大数。初期設定では、値は50に設定されています。設定した限界より多いアリルが存在する場合、アリルは入力サンプル中で存在する頻度により分類されます。最も一般的なアリルは出力されます。
--num-threads	使用するプロセッサースレッド数。初期設定は使用できるコア数です。
--gg-sites-list	ファイルの各部位において、gVCFジェノタイパーが深度情報を強制出力します。ファイルフォーマットはbedまたはbed.gzです。初期設定では、いずれのサンプルにも存在しない強制遺伝型部位は出力されません。これらの部位を表示するには、`--gg-discard-ac-zero`をfalseに設定します。
--gg-spvcf-out	出力を低密度プロジェクトVCFフォーマットで記述します。有効にするには、`true`に設定します。オプションは、初期設定で無効にされています。低密度プロジェクトVCFフォーマットの詳細については、『Sparse Project VCF: efficient encoding of population genotype matrices』を参照してください。[1]
--gg-enable-indexing	出力ファイルのtabixインデックスを構築します。オプションは初期設定で有効にされています。`--gg-output-format`をvcf.gzに設定して、`--gg-enable-indexing`を使用する必要があります。
--gg-drop-genotypes	バリアントごとにアリルのみを出力するように選択します。初期設定では、値はfalseに設定されています。`--gg-drop-genotypes`は、初期設定の出力でbcftools view -Gを実行するのと同等です。
--gg-write-phased-gt	`false`に設定すると、gVCFジェノタイパーは、入力ファイルのフェージング情報を無視します。ジェノタイパーは、サンプル中のフェージングされた遺伝型をフェージングされていないとして出力ファイルに記述します。このオプションは初期設定で有効にされています。
--gg-allele-list	**(オプション)**指定した部位での遺伝型を強制出力します。部位が含まれているvcf.gzまたはbcfファイルのパスを含める必要があります。
--gg-remove-nonref	**(オプション)**gVCFジェノタイパーの出力から<NON_REF>記号のアリルを除去します。このオプションは、<NON_REF>が付いたVCF行を処理できない下流のツールに対応するために使われます。

オプション	説明

--gg-sample-rename-mapfile	**(オプション)**結合した出力で名前を変更するサンプルのタブまたはカンマ区切りのマッピングを含むファイルへのパスを指定します。元のサンプル、ターゲット名、およびファイル名を含めるか、または/Perl regex pattern/,substitutionを使用します。以下は、ファイルの例です:
	`sampleX_3466,sampleX_3456`
	`sampleY_1234,sampleY_filtered_1234,sampleY_30x_filtered.gvcf.gz`
	`/(\w+)_badsuffix/ \1_bettersuffix`

--gg-concurrency-regions	**(オプション)**並列に処理する領域を指定します。BEDファイル、領域のカンマ区切りのリスト、または領域の行区切りのリストを指定できます。
--gg-discard-ac-zero	trueに設定した場合、gVCFジェノタイパーは、いずれのサンプルでもコールされていないバリアントアリル(hom-ref遺伝型)を出力しません。初期設定値はtrueです。

[1]Lin MF, Bai X, Salerno WJ, Reid JG. Sparse Project VCF: efficient encoding of population genotype matrices. *Bioinformatics*. 2020;36(22-23):5537-5538. doi:10.1093/bioinformatics/btaa1004

ジョイントジェノタイパーのオプション

マルチサンプルVCF、マルチサンプルgVCF、または1セットの単一サンプルgVCFから直接にジョイントジェノタイパーを実行できます。

ジョイントジェノタイパーでは、以下のパラメーターが使用できます。

オプション	説明
--enable-joint-genotyping	ジョイントジェノタイパーを実行するには、trueに設定します。
--output-directory	出力ディレクトリ。--output-directoryが必須です。
--output-file-prefix	すべての出力ファイルにラベル付けするのに使用されるプレフィックス。`--output-file-prefix`が必須です。
-r	ハッシュテーブルが存在するディレクトリ。

オプション	説明
--variant *--variant-list*	単一のgVCFファイルへのパスを指定します。複数の`--variant`オプションを使用して、複数のgVCFファイルを指定できます。最大200のgVCFをサポートしています。`--variant-list`を使用し、行当たり1つのバリアントファイルパスを使用して組み合わせる必要があるgVCFファイルのリストを含むファイルを指定します。
--pedigree-file	サンプル間の関係を記述するpedigreeファイルへのパスを指定します。詳細については、117 ページの「pedigreeモード」を参照してください。

ジョイント解析出力フォーマット

使用できるジョイント解析出力ファイルには、以下の2つがあります：

- **マルチサンプルVCF**：入力バリアントに応じて入力サンプルごとに遺伝型情報のある列を含むVCFファイル。
- **マルチサンプルgVCF**：単一サンプルのgVCFファイルがVCFファイルを拡張する方法と同様、マルチサンプルのVCFファイルの内容を拡張するgVCFファイル。バリアント部位間では、マルチサンプルgVCFには、各サンプルがリファレンスゲノムに対してホモ接合性が維持されている信頼度のレベルを記述する統計値が含まれています。マルチサンプルgVCFは、pedigreeまたは小さなコホートからの結果を単一のファイルに組み合わせるのに便利なフォーマットです。多数のサンプルを使用する場合、カバレッジの変動または入力サンプルのいずれかの変動により、新しいhom-refブロックが作成されます。これにより、大幅に断片化されたブロック構造および大きな出力ファイルを生成するため、作成の速度が低下する場合があります。

hom-refブロックのFORMATフィールド

hom-refブロックでは、以下のFORMATフィールドが一意的に計算されます。

- **FORMAT/DP**：値は、バンド内のすべての位置にわたる最小DPを表しています。
- **FORMAT/AD**：値は、DP = 中央値DPであるバンド内の位置を表しています。
- **FORMAT/AF**：値は、FORMAT/ADに基づいています。
- **FORMAT/PL**：値は、遺伝型仮説ごとのPhred尤度を表しています。hom-refブロックでは、FORMAT/PLの各値は、バンド内のすべての位置にわたる最小値を表しています。
- **FORMAT/SPL**および**FORMAT/ICNT**：hom-refブロックおよびバリアント記録の両方を含む、gVCF記録でレポートされるパラメーター。パラメーターを使用して、トリオの発端者でde *novo*であるバリアントの信頼度スコアを計算します。SNPでは、FORMAT/PLとFORMAT/SPLの両方が、De Novo Callerへの入力として使用されます。FORMAT/PLは、ジェノタイパーがコールされた場合、ジェノタイパーから得られたPhred尤度を表しています。FORMAT/SPLは、列方向の推定であるpregraphから得られたPhred尤度を表しています。FORMAT/SPLの各値は、バンド内のすべての位置にわたる最小値を表しています。INDELでは、gVCFファイルでレポートされるFORMAT/ICNTに基づいて、PL値がジョイントpedigreeコールステップで計算されます。FORMAT/ICNTは、2つの値で構成されています。最初の値はその位置でINDELを含まないリードの数で、2番目の値はその位置でINDELを含むリードの数です。FORMAT/ICNTの各値は、バンド内のすべての位置にわたる最大値を表しています。

以下の例のhom-refブロックでは、ICNTは、対象の位置で各サンプルにIndelが含まれているかどうかに関する情報を提供します。発端者にその位置でIndelが含まれており、親のICNTがIndelに対応するリードを示していない場合、発端者の信頼度スコアは高くなり、位置にIndel de *novo*コールがあります。

```
cchr1 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;FractionInformativeR
eads=0.600;R2_5P_bias=0.000
GT:AD:AF:DP:F1R2:F2R1:GQ:PL:SPL:ICNT:GP:PRI:SB:MB
0/1:28,32,0:0.533,0.000:60:20,21,0:8,11,0:15:73,0,12,307,157,464:255,0,25
5:23,10:3.8452e+01,1.3151e-
01,1.5275e+01,3.0757e+02,1.9173e+02,4.5000e+02:0.00,34.77,37.77,34.77,69.
54,37.77:4,24,7,25:8,20,14,18
```

SPLおよびICNT値はDRAGENに固有です。GATKバリアントコーラーは、SPLおよびICNT値を出力しません。

gVCFジェノタイパーでのgVCFフィールドの結合

gVCFファイルを結合する場合、gVCFジェノタイパーは、各入力ファイルからのgVCFフィールドをコピーして出力msVCFに組み合わせます。gVCFフィールドは、以下のようにして出力フィールドに組み合わせられます。

- **FORMAT/FT**：各入力ファイルからの`FILTER`フィールドを、各サンプルの出力フィールドにコピーします。
- **QUAL**：出力ファイル値は、各入力ファイルの`QUAL`値の合計です。
- **INFO/MQ**：出力ファイル値は、各入力フィールドの`INFO/MQ`値の合計であり、各入力フィールドの`INFO/DP`で重み付けされています。
- **INFO/MQRankSum**、**INFO/ReadPosRankSum**、および**INFO/R2_5P_bias**：出力ファイル値は、入力ファイルのフィールド値の中央値です。
- **INFO/DP**：出力ファイル値は、各入力ファイルの`INFO/DP`値の合計です。

FORMATフィールドは、各入力ファイルから出力ファイルにコピーされますが、出力ファイルにはすべてのサンプルには存在しないアリルが含まれているため、ジェノタイパーは欠落値を推定するか、またはプレースホルダー値を使用します。

ミトコンドリアバリアントコールを使用するgVCFジェノタイパー

ミトコンドリア（chrM）でのDRAGENバリアントコールは、常染色体バリアントコールとは以下の点が異なります。

- 遺伝型クオリティ（GQ）および遺伝型尤度（PL）メトリクスは、chrMでは計算されません。代わりに、DRAGENは、Phred値のクオリティスコアSQを使用します。DRAGEN v3.8またはそれ以前を使用する場合、LODが使用されます。
- 複対立遺伝子コールは、1行に結合されるのではなく、VCFファイルのいくつかの行に分割されます。各バリアントは、新しい行に置かれます。

de novoスモールバリアントフィルタリング

フィルタリングステップにより、ploidyが変更されている領域でのジョイントコールワークフローのde novoバリアントコールを同定します。de novoコールは、pedigreeメンバーの少なくとも1人が非二倍体遺伝型を示す領域の特異性を低減することができるため、de novoバリアントフィルタリングは、関連するバリアントをマーキングすることにより、コールセットの特異性を改善できます。

発端者のFORMAT/DNフィールドは、de novoバリアントがploidy変更SVまたはCNVとそれぞれ重複している場合、pedigreeの構造多型およびコピー数バリアントコールに基づいて、元のDeNovo値からDeNovoSVまたはDeNovoCNVに変更されます。その他すべてのバリアントの詳細は変更されないままで、入力VCFのすべてのバリアントもフィルタリングされた出力VCFに存在しています。逆位のような、ploidyが変化しない構造多型またはコピー数バリアントは、フィルタリングで考慮されません。例として、入力VCFのde novo SNVコールを示します

```
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;ReadPosRankSum=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
```

発端者、母親、または父親のSV重複との重複は、フィルタリングされた出力VCFで以下のように表されます：

```
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;ReadPosRankSum=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
```

以下は、ジョイントコールワークフローで返されるファイルに基づいて、de novoフィルタリングを実行するためのコマンドラインの例です：

```
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
```

de novoスモールバリアントフィルタリングのオプション

de novoバリアントフィルタリングでは、以下のオプションが使用されます：

オプション	説明
--dn-input-vcf	de novoコールステップからフィルタリングされるジョイントスモールバリアントVCF。
--dn-output-vcf	フィルタリングされたVCFを記述する必要があるファイル位置。指定していない場合、入力VCFが上書きされます。
--dn-sv-vcf	SVコールステップからのジョイント構造多型VCF。省略した場合、構造多型の重複の確認がスキップされます。
--dn-cnv-vcf	CNVコールステップからのジョイント構造多型VCF。省略した場合、コピー数バリアントの重複の確認がスキップされます。

生殖細胞系列バリアントスモールハードフィルタリング

DRAGENでは、VCF記録に存在するアノテーションに基づいて、VCF後のバリアントフィルタリングを提供します。初期設定および非初期設定バリアントハードフィルタリングについて、下で説明します。ただし、バリアントコーラーのコア内から相関誤差の仮説を組み込むという、DRAGENのアルゴリズムの性質により、パイプラインが真のバリアントをノイズと区別する際の機能を改善され、VCF後フィルタリングの依存性が大幅に低減されます。このため、DRAGENでの初期設定のVCF後フィルタリングは非常に単純です。

初期設定のスモールバリアントハードフィルタリング

生殖細胞系列パイプラインの初期設定のフィルターは以下のとおりです：

- ##FILTER=<ID=DRAGENSnpHardQUAL,Description="trueの場合は設定：QUAL < 10.41">
- ##FILTER=<ID=DRAGENIndelHardQUAL,Description="trueの場合は設定：QUAL < 7.83">
- ##FILTER=<ID=LowDepth,Description="trueの場合は設定：DP <= 1">
- ##FILTER=<ID=PloidyConflict,Description="染色体ploidyに一致していないバリアントコーラーからのジェノタイプコール">
- ##FILTER=<ID=base_quality,Description="この座位でのaltリードの中央値塩基クオリティが閾値を満たしていないためにフィルタリングされた部位">

- ##FILTER=<ID=lod_fstar,Description="バリアントが尤度の閾値を満たしていません（初期設定の閾値は6.3です）">
- DRAGENSnpHardQUALおよびDRAGENIndelHardQUAL：ミトコンドリアコンティグ以外のすべてのコンティグでは、初期設定のハードフィルタリングは、QUAL値の閾値処理のみで構成されています。別の初期設定のQUAL閾値は、SNPおよびIndelに適用されます。
- lod_fstar：ミトコンドリアコンティグでは、初期設定のハードフィルタリングは、LODスコアの閾値処理のみで構成されています。
- base_quality：ミトコンドリアコンティグでは、altリードの中央値塩基クオリティが閾値未満である部位にこのフィルターを適用します。
- LowDepth：このフィルターは、FORMAT/DP <= 1/のすべてのバリアントコールに適用されます。
- PloidyConflict：FemaleをDRAGENコマンドラインで指定するか、またはFemaleをPloidy Estimatorで検出する場合、Female被験者のchrYのすべてのバリアントコールにこのフィルターが適用されます。

非初期設定のスモールバリアントハードフィルタリング

VCF標準で説明されているように、DRAGENは、バリアントコールの基本的なフィルタリングをサポートしています。*--vc-hard-filter*オプションで任意の数のフィルターを適用できます。これは以下のように、数式のセミコロン区切りのリスト形式です：

<フィルターID>:<snp|indel|all>:<基準のリスト>、

ここで、基準のリストは数式のリスト自身であり、このフォーマットでは||（OR）演算子で区切られています：

<アノテーションID> <比較演算子> <値>

これらの数式要素の意味は以下のとおりです：

- **フィルターID**：フィルターの名前。該当する数式でフィルタリングされているコールのVCFファイルのFILTER列に入力されています。
- **snp/indel/all**：数式を適用する必要があるバリアントコールのサブセット。
- **アノテーションID**：フィルターの値を確認する必要があるバリアントコール記録アノテーション。サポートされているアノテーションは、FS、MQ、MQRankSum、QD、およびReadPosRankSumです。
- **比較演算子**：指定したフィルター値と比較するのに使用する数値比較演算子。サポートされている演算子は、<、≤、=、≠、≥、および>です。

例えば、以下の数式はラベル「SNP filter」FS < 2.1またはMQ < 100のSNPでマーキングされており、「indel filter」FS < 2.2またはMQ < 110の記録でマーキングされています：

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

この例は図示目的のみであり、DRAGEN V3出力で使用することは推奨されません。Illuminaでは、初期設定のハードフィルターを使用することを推奨しています。

値比較の組み合わせで唯一サポートされている操作はORであり、複数のアノテーションの算術的な組み合わせはサポートされていません。将来には、複雑な数式がサポートされる場合があります。

リードにわたる相関誤差のモデル化

DRAGENには、所定の集積でリードにわたる相関誤差をモデル化する2つのアルゴリズムが備えられています。

外部リード検出

外部リード検出（FRD）では、誤ってマップされたリードを検出します。FRDは確度計算を変更して、リードのサブセットが誤ってマップされた確率を考慮します。リードごとに独立してマッピングエラーが発生していると仮定するのではなく、FRDは、MAPQおよび偏ったAFのような裏付けを組み込むことにより、まとまった数のリードが誤ってマップされている確度を推定します。

通常マッピングエラーはまとまって発生しますが、マッピングエラーをリードごとに独立したエラーイベントとして取り扱うことにより、低MAPQや偏ったAFにもかかわらず、高い信頼度スコアを実現できます。信頼度スコアの過大評価を軽減するための1つの可能性のある戦略は、計算で使用されている最小MAPQに関する閾値を含めることです。ただし、この戦略は、裏付けを破棄して偽陽性を生成する場合があります。

FRDは、集積内のリードが外部リードの可能性がある（つまり、真の位置はリファレンスゲノム内のどこかである）という追加の仮説を組み込むことにより、従来のジェノタイピングアルゴリズムを拡張します。アルゴリズムは複数の特性（偏ったアリル頻度および低MAPQ）を利用して、この証拠を確度計算に組み込みます。

FNをレスキューして、遺伝型を修正し、リードをバリアントコーラーに入力するためのMAPQ閾値の低減を有効にすることにより、感度を改善します。FPを除去して遺伝型を修正することにより、特異性を改善します。

塩基クオリティ低下

塩基クオリティ低下（BQD）アルゴリズムにより、シーケンスシステムによって引き起こされる系統的で相関性のあるベースコールエラーを検出します。BQDは、これらのエラーの特定の特性（ストランドバイアス、リード内でのエラーの位置、塩基クオリティ）を利用して、アリルが真のバリアントではなく系統的なエラーイベントの結果である確度を推定します。

特定の座位で発生したエラーのまとまりには、それらを真のバリアントと区別する明確な特性があります。塩基クオリティ低下（BQD）アルゴリズムは、これらのエラーの特定の特性（ストランドバイアス、リード内でのエラーの位置、対象の座位でのリードの該当するサブセットにわたる低平均塩基クオリティ）を利用して、確度計算に組み込む検出メカニズムです。

方向バイアスフィルター

方向バイアスフィルターは、通常以下に関連するノイズを低減するように設計されています：

- ゲノムライブラリー調製時に導入されたアダプター前アーティファクト（例えば、熱、断片化、および金属汚染の組み合わせにより、シトシンまたはアデニンと対合の8-オキソグアニン塩基を生成する場合があり、最終的にはPCR増幅時にG→Tトランスバージョン変異を引き起こします）。
- FFPE（ホルマリン固定パラフィン包埋）アーティファクト。FFPEアーティファクトはシトシンのホルムアルデヒド脱アミノ化から生じており、CからTへの移行変異を引き起こします。

方向バイアスフィルターは、体細胞パイプラインでのみ使用できます。フィルターを有効にするには、*--vc-enable-orientation-bias-filter*オプションをtrueに設定します。初期設定はfalseです。

フィルタリングされるアーティファクトタイプは、*--vc-orientation-bias-filter-artifacts*オプションで指定できます。初期設定はC/T、G/Tですが、これはOxoGおよびFFPEアーティファクトに対応しています。有効な値は、C/T、G/T、またはC/T、G/T、C/Aです。

アーティファクト（または、アーティファクトとその逆相補）は2回リストできません。例えば、C→GおよびT→Aは逆相補であるため、C/T、G/Aは有効ではありません。

方向バイアスフィルターは、以下の情報を追加します。

- *##FORMAT=<ID=F1R2,Number=R,Type=Integer,Description=*"各アリルをサポートするF1R2ペア方向でのリードの数">
- *##FORMAT=<ID=F2R1,Number=R,Type=Integer,Description=*"各アリルをサポートするF2R1ペア方向でのリードの数">
- *##FORMAT=<ID=OBC,Number=1,Type=String,Description=*"方向バイアスフィルターの塩基コンテキスト">
- *##FORMAT=<ID=OBPa,Number=1,Type=String,Description=*"アーティファクトの方向バイアスプライア">
- *##FORMAT=<ID=OBParc,Number=1,Type=String,Description=*"逆相補アーティファクトの方向バイアスプライア">
- *##FORMAT=<ID=OBPsnp,Number=1,Type=String,Description=*"実際のバリアントの方向バイアスプライア">

dbSNPアノテーション

生殖細胞系列、Tumor-Normal体細胞、またはTumor-Only体細胞モードでは、DRAGENはdbSNPデータベースでバリアントコールを検索して、検出した一致バリアントに対してアノテーションを追加できます。dbSNPデータベース検索を有効にするには、*--dbsnp*オプションをdbSNPデータベースVCFまたはvcf.gzファイルへのフルパスに設定しますが、これはリファレンスの順序でソートする必要があります。

出力VCFの各バリアントコールでは、コールがCHROM、POS、REF、および少なくとも1つのALTのデータベースエントリーに一致した場合、一致したデータベースエントリーのrsIDが出力VCFの該当するコールのID列にコピーされます。さらに、DRAGENは、データベースで検出されたコールのINFOフィールドにDBアノテーションを追加します。

DRAGENは、リファレンスシーケンス/コンティグの名前に基づいてバリアントコールを照合しますが、dbSNPを構築するのに使用されるリファレンスが、アライメントおよびバリアントコールで使用されるリファレンスと同じであると断言する別の方法はありません。選択したアノテーションデータベースのコンティグが、アライメント/バリアントコールリファレンスのコンティグに一致していることを確認します。

VCFファイルの自動生成MD5SUM

VCF出力ファイルのMD5SUMファイルは、自動的に生成されます。このファイルは同じ出力ディレクトリに置かれており、VCF出力ファイルと同じ名前ですが、.md5sum拡張子が最後に付加されています。例えば、whole_genome_run_123.vcf.md5sumです。MD5SUMファイルは1行のテキストファイルであり、VCF出力ファイルのmd5sumが含まれています。このmd5sumは、Linux md5sumコマンドの出力に厳密に一致しています。

強制ジェノタイピング

DRAGENは、スモールバリアントコールの強制ジェノタイピング（ForceGT）をサポートしています。ForceGTを使用するには、強制ジェノタイピングするスモールバリアントのリストで*--vc-forcegt-vcf*オプションを使用します。スモールバリアントの入力リストは、*.vcfまたは*.vcf.gzファイルにすることができます。

ForceGTの現在の制限事項は以下のとおりです：

- ForceGTは、V3モードの生殖細胞系列スモールバリアントコールでサポートされています。V1、V2、およびV2+モードはサポートされていません。
- ForceGTは、体細胞スモールバリアントコールでもサポートされています。
- ForceGTは、ジョイントジェノタイピングを通して伝播しません。

ForceGT入力

DRAGENは単一のForceGT VCF入力ファイルのみをサポートしていますが、このファイルは以下の要件を満たす必要があります：

- バリアントコールで使用されるVCFと同じリファレンスコンティグを有している。
- リファレンスコンティグの名前と位置でソートされる。
- ノーマライズされている（簡潔で左揃え）。
- 複雑なバリアントが含まれていない（refアリルからALTアリルに移動するために、複数の置換/挿入/欠失が必要なバリアント）。例えば、未定義の動作で以下の結果に類似しているForceGT VCFのバリアントです：

```
chrX 153592402 GTTGGGGATGCTGAC CACCCTGAAGGG
```

以下のノーマライズされていないバリアントは、DRAGENで未定義の動作を引き起こします：

- 簡潔ではない：`chrX 153592402 GC GCG`
- 簡潔な表現：`chrX 153592403 C CG`

ForceGT操作と予測される結果

ForceGTでスモールバリアントコールを実行すると、DRAGENは、コマンドラインの入力としてForceGT VCFを使用することによりVCFを生成します。VCF出力ファイルには、以下のようにすべての通常のコールとForceGTコールが含まれています：

- バリアントコーラーでコールされなかったForceGTコールでは（一般的ではない）、コールはINFOフィールドにおいてFGTでタグ付けされます。
- バリアントコーラーでコールされ、フィルターフィールドがPASSの生殖細胞系列ForceGTコールでは（一般的）、コールはINFOフィールドにおいてNML:FGTでタグ付けされます（NMLは正常を示します）。体細胞モードでは、コールはSOM;FGTでタグ付けされます。
- バリアントコーラーによる正常コール（およびPASS）で、ForceGTコールがない場合（正常）、余分なタグは追加されません（NMLタグなし、FGTタグなし）。

このスキームは、FGTのみ、ForceGT入力と正常コールの両方で一般的、および正常コールであることが原因で存在するコールを区別します。

入力ForceGT VCFのすべてのバリアントは遺伝型で処理され、出力VCFファイルに存在しています。以下の表は、バリアントでレポートされるGTを示しています。

条件	レポートされるGT
カバレッジのない位置	./.または.|.
カバレッジはあるがALTアリルをサポートしているリードがない位置	0/0または0|0
カバレッジがありALTアリルをサポートしているリードがある位置	パイプラインに依存(生殖細胞系列/体細胞)

DRAGENが、入力ForceGT VCFで指定したものとは異なるバリアントをコールした場合、出力VCFには同じ位置に以下の複数のエントリーが含まれています：

- 初期設定DRAGENバリアントコールの1つのエントリー
- 該当する位置で入力ForceGT VCFに存在するすべてのバリアントコールそれぞれに対する1つのエントリー。

```
chrX 100 G C [Default DRAGEN variant call]
chrX 100 G A [Variant in ForceGT vcf]
```

ターゲットBEDファイルが入力ForceGT VCFとともに指定されている場合、出力gVCFファイルには、BEDファイル位置と重複するForceGTバリアントのみが含まれています。

コピー数バリアントコール

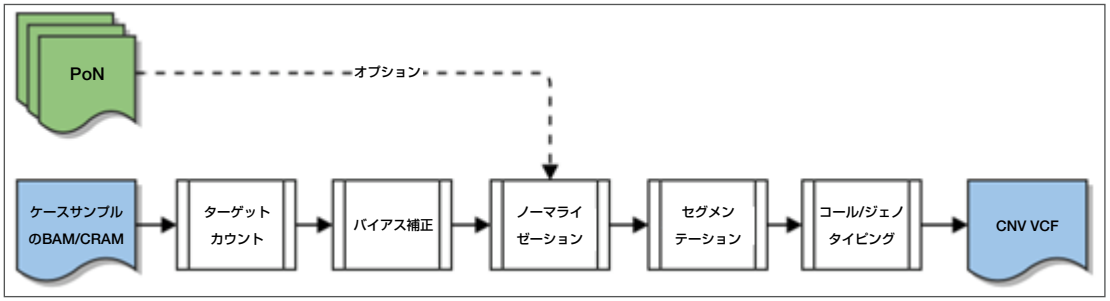
DRAGENコピー数バリアント（CNV）パイプラインでは、次世代シーケンサー（NGS）データを使用してCNVイベントをコールできます。このパイプラインでは、生殖細胞系列解析での全ゲノムシーケンス（WGS）データおよび全エクソームシーケンス（WES）データの処理を含む、DRAGENホストソフトウェアを介した単一インターフェースで複数のアプリケーションをサポートしています。

DRAGEN CNVパイプラインでは、2つのノーマライゼーション動作モードをサポートしています。2つのモードで異なるノーマライゼーション手法を適用し、アプリケーション、例えば、WGS対WESに基づいて異なるバイアスを処理します。初期設定のオプション設定では、速度と精度の点から最適のトレードオフの提供を試行していますが、独自のワークフローでは細かくチューニングしたオプション設定が必要になる場合があります。

CNVワークフロー

DRAGEN CNVパイプラインは、以下の図に示されているワークフローに従っています。

図 2　DRAGEN CNVパイプラインワークフロー



DRAGEN CNVパイプラインでは、ハードウェアアクセラレーションや十分なI/O処理のような、他のパイプラインで使用できるDRAGENプラットフォームの多数の特徴を使用します。DRAGENホストソフトウェアでCNV処理を有効にするには、*--enable-cnv*コマンドラインオプションをtrueに設定します。

CNVパイプラインには、以下の処理モジュールが備えられています：

- **ターゲットカウント**：アライメントからのリードカウントおよびその他のシグナルのビニング。
- **バイアス補正**：内因性のシステムバイアスの補正。
- **ノーマライゼーション**：正常なploidyレベルの検出およびケースサンプルのノーマライゼーション。
- **セグメンテーション**：ノーマライズしたシグナルのセグメンテーションを介したブレークポイント検出。
- **コール/ジェノタイピング**：閾値処理、スコアリング、資格付与、および推定イベントのコピー数バリアントとしてのフィルタリング。

ノーマライゼーションモジュールは、必要に応じて正常サンプルのパネル（PoN）を取り込むことができます。これは、コホートまたは集団サンプルが即座に使用できる際に使用されます。その他のモジュールはすべて、異なるCNVアルゴリズム間で共有されています。

シグナルフロー解析

以下の図は、シグナルがさまざまなステージを通して横断する際のDRAGEN CNVパイプラインのステップの概要を示しています。これらの図は例であり、DRAGEN CNVパイプラインから生成されるプロットとは同一ではありません。

DRAGEN CNVパイプラインでの最初のステップは、ターゲットカウントステージです。ターゲットカウントステージでは、リードカウントや不適当なペアのようなシグナルを抽出して、そのシグナルをターゲット間隔に入力します。

図 3　リードカウントシグナル

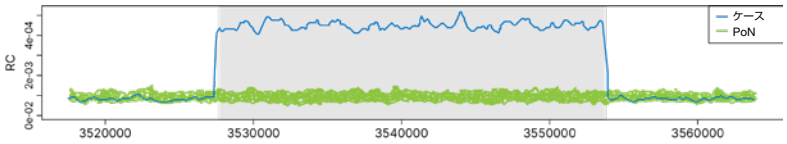
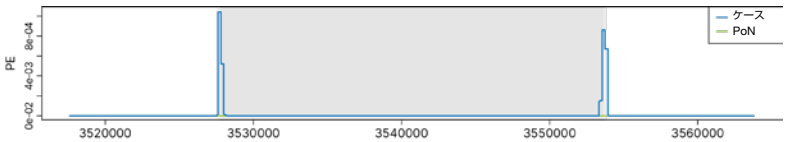
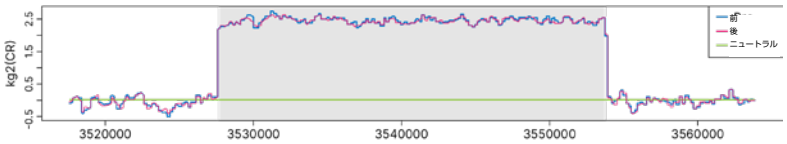


図 4　不適当なペアシグナル



次に、ケースサンプルを正常サンプルのパネルまたは推定した正常なploidyレベルに対してノーマライズします。その他のバイアスはすべてシグナルから減算して、イベントレベルのシグナルを増幅します。

図 5　タンジェントノーマライゼーション前/後



ノーマライズしたシグナルを、使用できるセグメンテーションアルゴリズムの1つを使用してセグメント化します。イベントがセグメントからコールされます。

図 6　セグメント

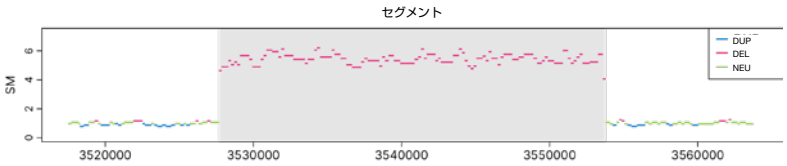
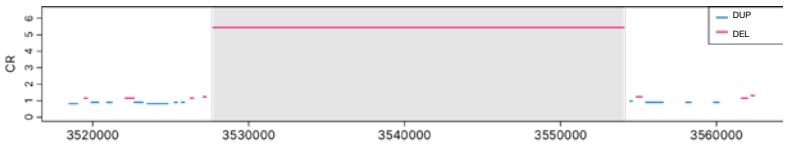


図 7　コールされたイベント



出力VCFで、イベントが得点されて放出されます。

CNVパイプラインのオプション

以下は、CNVパイプラインを制御するためにDRAGENホストソフトウェアと共有される最上位のオプションです。BAMまたはCRAMファイルは、CNVパイプラインに入力できます。DRAGENのマッパーとアライナーを使用する場合は、FASTQファイルを使用できます。

オプション	説明
--bam-input	処理対象のBAMファイル。
--cram-input	処理対象のCRAMファイル。
--enable-cnv	CNV処理を有効または無効にします。CNV処理を有効にするには、trueに設定します。
--enable-map-align	マッパーおよびアライナーモジュールを有効にします。初期設定はtrueであるため、このオプションをfalseに設定しない限り、すべての入力リードが再マップされてアライメントされます。
--fastq-file1 *--fastq-file2*	処理対象のFASTQファイル。
--output-directory	すべての結果が格納される出力ディレクトリ。
--output-file-prefix	すべての結果ファイル名の前に付加される出力ファイルの接頭辞。
--ref-dir	DRAGENリファレンスゲノムのハッシュテーブルディレクトリ。

CNVパイプライン入力

DRAGEN CNVパイプラインでは、複数の入力フォーマットをサポートしています。最も一般的なフォーマットは、すでにマップされてアライメントされているBAMまたはCRAMファイルです。まだマップもアライメントもされていないデータがある場合は、137 ページの 「アライメントファイルの生成」 を参照してください。

FASTQ入力を使用してBAMおよびCRAMファイルを生成せずにDRAGEN CNVパイプラインを直接実行する際には、DRAGENマップ/アライメントステージから直接アライメント記録をストリーミングする手順について138 ページの 「アライメントのストリーミング」 を参照してください。

リファレンスハッシュテーブル

DRAGEN CNVパイプラインでは、その他のパイプラインで必要なその他のオプションに加えて、*--enable-cnv*オプションをtrueに設定してハッシュテーブルを生成する必要があります。*--enable-cnv*がtrueの場合、DRAGENは、CNVアルゴリズムがマップ可能性バイアスを抑制するのに使用するk-mer特異性マップを追加で生成します。リファレンスハッシュテーブルごとに一度、k-mer特異性マップファイルを生成する必要があります。生成には、全ヒトゲノム当たり約1.5時間を要します。

リファレンスハッシュテーブルは、リファレンスゲノムの事前生成済みバイナリー表現です。ハッシュテーブルの生成の詳細については、10 ページの 「リファレンスゲノムの準備」 を参照してください。

以下は、コマンドがハッシュテーブルを生成する例です。

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

アライメントファイルの生成

以下のコマンドラインの例では、入力タイプに応じてDRAGENのマップ/アライメントパイプラインを実行する方法について示します。マップ/アライメントパイプラインでは、パイプラインで使用できるBAMまたはCRAMファイルの形式でアライメントファイルを生成します。

まだマップもアライメントもされていないすべてのサンプルのアライメントファイルを生成する必要があります。各サンプルには、一意のサンプル識別子が備えられている必要があります。*--RGSM*オプションを使用して、識別子を指定します。BAMおよびCRAM入力ファイルでは、サンプル識別子がファイルから取得されるため、*--RGSM*オプションは必要ありません。

以下のコマンドは、FASTQファイルをマップしてアライメントする例です：

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

以下のコマンドは、既存のBAMファイルをマップしてアライメントする例です：

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

以下のコマンドは、既存のCRAMファイルをマップしてアライメントする例です：

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

アライメントのストリーミング

DRAGENはFASTQサンプルをマップしてアライメントしてから、CNVコーラーやハプロタイプバリアントコーラーのような下流のコーラーに対してそのサンプルを直接ストリーミングできます。このプロセスを使用してBAMまたはCRAMファイルの生成をスキップできますが、これにより追加ファイルの格納の必要性を回避します。

アライメントを直接DRAGEN CNVパイプラインにストリーミングするには、通常のDRAGENマップ/アライメントワークフローを通してFASTQサンプルを実行してから、引数を追加してCNVを有効にします。以下は、コマンドラインがFASTQファイルをマップしてアライメントしてから、ファイルを生殖細胞系列CNV WGSパイプラインに送る例です。

```
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
```

CNVとハプロタイプバリアントコーラーを同時に実行する方法については、158 ページの 「同時CNVおよびスモールバリアントコール」 を参照してください。

ターゲットカウント

DRAGEN CNVパイプラインでの最初の処理ステージは、ターゲットカウントステージです。このステージは、アライメントを間隔ごとにビニングします。CNV処理の一次解析フォーマットはターゲットカウントファイルであり、ここには、下流の処理で使用されるアライメントから抽出された機能シグナルが含まれています。ビニング戦略、間隔サイズ、およびその境界は、ターゲットカウント生成オプションおよび使用するノーマライゼーション手法により制御されます。

全ゲノムシーケンスデータを処理する際には、リファレンスハッシュテーブルから間隔が自動生成されます。ビニングでは、リファレンスハッシュテーブルからの一時コンティグのみを考慮します。*--cnv-skip-contig-list*オプションで、回避する追加のコンティグを指定できます。

全エクソームシーケンスデータでは、DRAGENは*--cnv-target-bed*オプションで指定されたターゲットBEDファイルを使用して、解析の間隔を決定します。

ターゲットカウントステージでは、`.target.counts.gz`ファイルを生成します。後にBAMまたはCRAMの代わりにファイルを使用するには、ノーマライゼーションステージにおいて*--cnv-input*オプションでファイルを指定します。`.target.counts.gz`ファイルは、DRAGEN CNVパイプラインの中間ファイルであり、変更してはいけません。

`.target.counts.gz`ファイルは、以下の列を含むタブ区切りの圧縮テキストファイルです：

- コンティグ識別子
- 開始位置
- 終了位置
- ターゲット間隔名
- この間隔でのアライメント数
- この間隔で不適当に対合されているアライメント数

*.target.counts.gzファイルの例を以下に示します。

```
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
```

全ゲノム

サンプルが全ゲノムである場合、*--cnv- interval-width*オプションで効果的なターゲット間隔幅を指定します。サンプルのカバレッジが高くなるほど、検出できる解像度が高くなります。すべてのサンプルが一致した間隔を備えている必要があるため、このオプションは正常サンプルのパネルで実行する際に重要になります。セルフノーマライゼーションでは、効果的な幅は指定した値より大きくなる場合があります。

WGSの初期設定値は1000 bpで、サンプルカバレッジ≥ 30xです。

サンプル当たりのWGSカバレッジ	推奨される解像度*（bp）
5	10000
10	5000
≥ 30	1000

*WGS解析で`cnv-interval-width` ≤ 250 bpを使用すると、ランタイムが劇的に増大する場合があります。

リファレンスのすべての一次コンティグに対して、間隔が自動生成されます。DRAGENでは、USCSまたはGRC規則があるリファレンスのみをサポートしています。例えば、`chr1, chr2, chr3, ..., chrX, chrY or 1, 2, 3, ..., X, Y`です。スキップするコンティグのリストを指定するには、*--cnv-skip-contig-list*オプションを使用します。このオプションでは、コンティグ識別子のカンマ区切りのリストを取得します。コンティグ識別子は、使用しているリファレンスハッシュテーブルに一致している必要があります。初期設定では、ミトコンドリア染色体のみがスキップされます。一次ではないコンティグは処理されません。

例えば、染色体M、X、およびYをスキップするには、以下のオプションを使用します：

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

全エクソーム

サンプルが全エクソームサンプルである場合、*--cnv-target-bed $TARGET_BED*オプションでターゲットBEDファイルを指定します。

ターゲットBEDファイルの間隔は、ターゲットキャプチャーキットに基づいてアライメントが予測される領域を示しています。BEDファイルの間隔は、*cnv-interval-width*の値に応じて、さらに小さなサイズの間隔に分割されます。

標準のBEDファイルを使用するには、ファイルにヘッダーが存在していないことを確認します。この場合、DRAGENバリアントコーラーの操作と同様に、3番目の列より後の列はすべて無視されます。

ターゲットカウントオプション

以下のオプションにより、ターゲットカウントの生成を制御します。

オプション	説明
--cnv-counts-method	ターゲットビンでカウントするアライメントのカウントメソッドを指定します。値は、midpoint、start、またはoverlapです。正常サンプルのパネルアプローチを使用する際の初期設定値はoverlapですが、これはアライメントがターゲットビンのいずれかの部分と重複する場合、そのビンでアライメントがカウントされるという意味です。セルフノーマライゼーションモードでは、初期設定のカウントメソッドはstartです。
--cnv-min-mapq	ターゲットカウント生成時にカウントするアライメントの最小MAPQを指定します。初期設定値は、セルフノーマライゼーションでは3、それ以外では20です。正常サンプルのパネルのカウントを生成する際には、すべてのMAPQ0アライメントがカウントされます。
--cnv-target-bed	サンプルカバレッジに対するターゲット間隔を示す、適切にフォーマットされたBEDファイルを指定します。WES解析で使用します。
--cnv-interval-width	CNV処理でのサンプリング間隔の幅を指定します。このオプションでは、効果的なウィンドウサイズを制御します。初期設定は、WGS解析では1000、WES解析では500です。

オプション	説明
--cnv-skip-contig-list	WGS解析の間隔を生成する際にスキップするコンティグ識別子のカンマ区切りのリストを指定します。指定していない場合、スキップされる初期設定のコンティグは、`chrM,MT,m,chrm`です。

ターゲットカウント欠落領域

所定のリファレンスでBEDファイルが指定されていないWGSの場合、毎回同じ間隔を生成する必要があります。作成した間隔では、ハッシュテーブル生成時に作成したk-mer特異性マップを使用してリファレンスゲノムのマップ可能性を考慮します。欠落領域は複雑な領域であり、解析からの間隔欠落でアライメントおよび結果をカウントしません。欠落領域は、セントロメア、テロメア、および複雑性の低い領域です。隣接領域に十分なシグナルが存在する場合、領域でアライメントカウントが発生しなくても、イベントは依然としてこれらの欠落領域に広がることができます。イベントは、セグメンテーションステージで処理されます。

GCバイアス補正

GCバイアスでは、ゲノムにわたるGCコンテンツとリードカバレッジ間の関係を測定します。バイアスは、ライブラリー調製、キャプチャーキット、シーケンスシステム差、およびマッピングで発生する場合があります。バイアスにより、CNVイベントをコールするのが困難になる場合があります。DRAGEN GCバイアス補正モジュールでは、これらのバイアスの修正を試みます。

GCバイアス補正モジュールは即座にターゲットカウントステージに従い、`*.target.counts.gz`ファイルで動作します。GCバイアス補正により、ファイルのGCバイアス補正バージョンを生成しますが、これにはファイル名に`*.target.counts.gc-corrected.gz`拡張子があります。GCバイアス補正バージョンは、WGSデータを処理する際の下流の処理で推奨されます。WESでは、十分なターゲット領域が存在する場合、GCバイアス補正カウントも使用できます。

代表的なキャプチャーキットでは、対象の領域にわたって200,000を超えるターゲットがあります。BEDファイルに200,000より少ないターゲットが存在するか、またはターゲット領域がゲノムの特定の領域に対してローカライズされている場合（GCバイアス統計値が歪められている場合がある）、GCバイアス補正を無効にする必要があります。

以下のオプションで、GCバイアス補正モジュールを制御します。

オプション	説明
--cnv-enable-gcbias-correction	ターゲットカウントの生成時に、GCバイアス補正を有効または無効にします。初期設定はtrueです。
--cnv-enable-gcbias-smoothing	指数カーネル関数のある隣接のGCビンにわたるGCバイアス補正の平滑化を有効または無効にします。初期設定はtrueです。
--cnv-num-gc-bins	GCバイアス補正のビン数を指定します。各ビンは、GCコンテンツの割合を表しています。許可されている値は、10、20、25、50、または100です。初期設定は25です。

ノーマライゼーション

DRAGEN CNVパイプラインでは、2つのノーマライゼーションアルゴリズムをサポートしています：

- **セルフノーマライゼーション**：解析下のサンプルからの常染色体二倍体レベルを推定して、ノーマライズするベースラインレベルを決定します。性染色体およびPAR領域は、サンプルの性に基づいて処理されます。
- **正常サンプルのパネル**：CNVイベントをコールするベースラインレベルを決定するために、追加の一致した正常サンプルを使用するリファレンスベースのノーマライゼーションアルゴリズム。この場合の一致した正常サンプルとは、ケースサンプルと同じライブラリー調製およびシーケンスワークフローを経たものであることを意味します。

使用するアルゴリズムは、使用できるデータとアプリケーションに依存しています。以下のガイドラインを使用して、ノーマライゼーションのモードを選択します。

セルフノーマライゼーション

- 全ゲノムシーケンス
- 単一サンプル解析
- 追加の一致したサンプルは即座には使用できません
- 単一の呼び出しを介した単純なワークフロー
- `chr1, chr2, chr3, ..., chrX, chrY or 1, 2, 3, ..., X, Y`命名規則を使用しているリファレンスのみがサポートされます

正常サンプルのパネル

- 全ゲノムシーケンス
- 全エクソームシーケンス
- 体細胞パネルを含む、ターゲットパネル
- 追加の一致したサンプルが使用できます
- 非ヒトサンプル

セルフノーマライゼーション

DRAGEN CNVパイプラインでは、標準サンプルおよび正常サンプルのパネルを必要としないセルフノーマライゼーションモードを提供します。このモードを有効にするには、*--cnv-enable-self-normalization*をtrueに設定します。セルフノーマライゼーションモードでは2つのステージを実行する必要性を回避して、時間を節約できます。また、ケースサンプル内の統計値を使用して、コールするためのベースラインを決定します。

セルフノーマライゼーションではケースサンプル内の統計値を使用するため、データが不十分になる可能性があることにより、WESおよびターゲットシーケンスではこのモードは推奨されません。

セルフノーマライゼーションモードは、全ゲノムシーケンスの単一サンプル処理で推奨されるアプローチです。パイプラインはセグメンテーションおよびコールステージまで継続し、最終コールイベントを生成します。

```
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
```

FASTQサンプルから実行する場合、操作の初期設定モードはセルフノーマライゼーションです。

セルフノーマライゼーションモードで操作する場合は、ユニークなk-mer位置数に基づいて、ターゲットカウントステージ時に使用される*--cnv-interval-width*オプションが効果的な間隔幅になります。通常はこのオプションを変更する必要はありません。

セルフノーマライゼーションは、リファレンスゲノムに基づいて解析時に使用するターゲット間隔を自動生成し、標準のヒトリファレンスにのみ対応しています。非標準ヒトリファレンスでは、処理するBEDファイルおよび正常サンプルのパネルアプローチが必要です。

正常サンプルのパネル

正常サンプルのパネルモードでは1セットの一致した正常サンプルを使用して、CNVイベントをコールするベースラインレベルを決定します。これらの一致した正常サンプルは、ケースサンプルで使用したのと同じライブラリー調製およびシーケンスワークフローから得る必要があります。これにより、アルゴリズムでサンプル特異的ではないシステムレベルバイアスを減算します。

このモードでは、DRAGEN CNVパイプラインは2つの別個のステージに細分されます。各サンプル、ケース、正常サンプルでターゲットカウントステージを実行し、アライメントをビニングします。次に、正常サンプルのパネルに対してケースサンプルでノーマライゼーションおよびコール検出ステージを実行し、イベントを決定します。

ターゲットカウントステージ

サンプルがリファレンスとして使用されるか、またはサンプルが解析下のケースサンプルであるかにかかわらず、すべてのサンプルでターゲットカウントを実行します。ケースサンプルおよび正常サンプルのパネルとして使用されるすべてのサンプルは同一の間隔を持つ必要があるため、同一の設定で生成する必要があります。ターゲットカウントステージでは、GCバイアス補正も実行します。GCバイアス補正は、初期設定で有効にされています。

以下の例は、WGS処理に対するものです。エクソーム処理については、140 ページの 「全エクソーム」を参照してください。

以下は、BAMファイルを処理するコマンドの例です。

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

以下は、CRAMファイルを処理するコマンドの例です。

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

ノーマライゼーションおよびコール検出ステージ

正常サンプルのパネルを使用する際のCNVパイプラインでの次のステップは、ノーマライゼーションを実行してコールすることです。このステップには正常サンプルのパネルの選択が含まれていますが、これはリファレンスベースの中央値ノーマライゼーションで使用されるターゲットカウントファイルのリストです。

他のワークフローの組み合わせで解析を実行できます。CNVイベントは、使用するリファレンスサンプルでコールされます。理想的なのは、正常サンプルのパネルが、解析下のケースサンプルのワークフローと同一のライブラリー調製およびシーケンスワークフローに従うことです。性染色体のコールでは、正常サンプルのパネルに、雄および雌サンプルの両方を均衡させて含める必要があります。DRAGENは、パネルの各サンプルの予測される性に基づいて、性染色体のコールを自動的に処理します。

最適のバイアス補正では、パネルとして最小50サンプルを推奨します。DRAGENは単一サンプルパネルで実行できますが、単一サンプルパネルは、パネルサンプルでコピー数が変化している検査サンプルで人為的にコールする場合があります。

正常サンプルのパネル（PON）を生成するには、ファイルの各行に、ターゲットカウントステージから生成された`.target.counts.gz`ファイルを指し示すパスが含まれているプレーンテキストファイルを作成します。パスが現在の作業ディレクトリに相対的である場合、相対パスがサポートされます。後でワークフローを使用するか、または他のユーザーと共有する場合は、絶対パスを推奨します。

以下は、ターゲットカウントステージからのGC修正ファイルのサブセットを使用する、PONファイルの例です。

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

または、*--cnv-normals-file*オプションで、正常サンプルのパネルで使用されるファイルを指定できます。このオプションでは単一のファイル名を使用しており、複数回指定できます。

PONファイルを作成した後、*--cnv-input*オプションでケースサンプルを指定し、*--cnv-normals-list*オプションでPONファイルを指定することにより、コーラーを実行できます。前のステージでGCバイアス補正カウントを使用している場合、GCバイアス補正を再度実行する必要はありません。GCバイアス補正を無効にするには、*--cnv-enable-gcbias-correction*をfalseに設定します。

例えば、以下のコマンドでは、正常サンプルのパネルに対してケースサンプルをノーマライズしています。

```
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
```

ノーマライゼーションオプション

これらのオプションでは、正常サンプルのパネルのプレコンディショニングおよびケースサンプルのノーマライゼーションを制御します。

オプション	説明
--cnv-enable-self-normalization	セルフノーマライゼーションモードを有効/無効にしますが、正常サンプルのパネルは必要ありません。
--cnv-extreme-percentile	サンプルをフィルタリングで除外する際の極端な中央値パーセンタイル値を指定します。初期設定は2.5です。
--cnv-input	正常サンプルのパネルを使用する際に、解析下のケースサンプルのターゲットカウントファイルを指定します。

オプション	説明
--cnv-normals-file	正常サンプルのパネルで使用される`target.counts.gz`ファイルを指定します。このオプションは、ファイルごとに複数回、1回使用できます。
--cnv-normals-list	正常サンプルのパネルとして使用されるリファレンスターゲットカウントファイルのリストへのパスを含むテキストファイルを指定します。後でワークフローを使用するか、または他のユーザーと共有する場合は、絶対パスを推奨します。パスが現在の作業ディレクトリに相対的である場合、相対パスがサポートされます。
--cnv-max-percent-zero-samples	ターゲットで許可されているゼロカバレッジサンプル数を指定します。ターゲットが指定した閾値を超えている場合、ターゲットはフィルタリングで除外されます。初期設定値は5%です。オプションは、使用される正常サンプル数にセンシティブです。それに応じて閾値を確実に調整します。正常サンプルのパネルのサイズが小さく、閾値が調整されていない場合、オプションにより目的ではなかったターゲットをフィルタリングして除外できます。
--cnv-max-percent-zero-targets	サンプルで許可されているゼロカバレッジターゲット数を指定します。サンプルが指定した閾値を超えている場合、サンプルはフィルタリングで除外されます。初期設定値は2.5%です。オプションは、ターゲット間隔の合計数にセンシティブです。それに応じて閾値を確実に調整します。キャプチャーキットのプローブ数が小さく、閾値が調整されていない場合、オプションにより目的ではなかったターゲットをフィルタリングして除外できます。
--cnv-target-factor-threshold	使用できるターゲットをフィルタリングして除外するための正常サンプルのパネル中央値の下部パーセンタイルを指定します。初期設定は、全ゲノム処理では1%で、ターゲットシーケンス処理では5%です。
--cnv-truncate-threshold	極端な外れ値を切り捨てるためのパーセンテージ閾値を指定します。初期設定は0.1%です。

セグメンテーション

ケースサンプルがノーマライズされた後、サンプルはセグメンテーションステージを通過します。DRAGENは、以下のアルゴリズムを含む、複数のセグメンテーションアルゴリズムを実装しています：

- 円形バイナリーセグメンテーション（CBS）
- レベルシフトモデル（SLM）

SLMアルゴリズムには、SLM、不均質なSLM（HSLM）、適応可能なSLM（ASLM）という3つのバリアントがあります。HSLMはエクソーム解析に使用され、等間隔ではないターゲットキャプチャーキットを処理します。ASLMには、コピー数の変更とは異なり、カバレッジ深度の技術多様性の追加のサンプル独自の推定が含まれています。推定は、固定ウィンドウまたはbアリルの割合に基づいた事前セットのセグメント内の中央値分散に基づいています。ASLMアルゴリズムは、ノイズの多いまたは波状のサンプルが原因の過大セグメンテーションを軽減します。

初期設定では、SLMは生殖細胞系列全ゲノム処理のセグメンテーションアルゴリズムで、ASLMは体細胞全ゲノム処理のアルゴリズムであり、HSLMは全エクソーム処理のアルゴリズムです。

ターゲットシーケンスワークフローでは、`--cnv-segmentation-bed`でも実行できます。オプションでは、コピー数を推定するセグメントを事前定義して、ワークフローのセグメンテーションステップをスキップします。149 ページの 「ターゲットセグメンテーション（セグメントBED）」 を参照してください。

オプション	説明
--cnv-segmentation-mode	実行するセグメンテーションアルゴリズムを指定します。以下の値が使用できます: • `bed` • `cbs` • `slm`：生殖細胞系列WGS解析の初期設定。 • `aslm`：体細胞WGS解析の初期設定。 • `hslm`：ターゲット/WES解析の初期設定。
--cnv-merge-distance	結合できる2つのセグメント間の塩基対の最大数を指定します。WGSでの初期設定値は0ですが、これはセグメントを直接隣接にする必要があるという意味です。WES解析では、ターゲット間隔のあけかたにより、このパラメーターは初期設定で無効にされています。
--cnv-merge-threshold	2つの隣接セグメントを結合する必要がある最大セグメントの平均値差を指定します。セグメント平均値は、線形コピー割合値として表されています。初期設定はWGSでは0.2で、WESでは0.4です。結合を無効にするには、値を0に設定します。

円形バイナリーセグメンテーション

円形バイナリーセグメンテーションはDRAGENに直接実装されており、アレイCGHデータ解析の高速円形バイナリーセグメンテーション[1]を基にして、NGSデータの感度を改善するように拡張されています。

以下のオプションにより、円形バイナリーセグメンテーションを制御します。

オプション	説明
--c-alpha	検定で変更点を受け入れる有意性レベルを指定します。初期設定は0.01です。

オプション	説明
--cnv-cbs-eta	並べ替えメソッドを使用する際の早期停止で、連続した境界のタイプIエラー率を指定します。初期設定値は0.05です。
--cnv-cbs-kmax	並べ替えの小さなセグメントの最大幅を指定します。初期設定は25です。
--cnv-cbs-min-width	変更後のセグメントのマーカーの最小数を指定します。初期設定は2です。
--cnv-cbs-nmin	最大統計近似のデータの最小長を指定します。初期設定は200です。
--cnv-cbs-nperm	P値計算で使用される並べ替え数を指定します。初期設定は10000です。
--cnv-cbs-trim	分散計算でトリミングされるデータの割合を指定します。初期設定は0.025です。

[1]Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics. 2007;23(6):657-663. doi:10.1093/bioinformatics/btl646

レベルシフトモデルセグメンテーション

レベルシフトモデル（SLM）セグメンテーションモードは、SLMSuite：ゲノムプロファイルのセグメント化のためのアルゴリズムのスイート[2]のR実装から得られます。

オプション	説明
--cnv-slm-eta	平均値プロセスが値を変更するベースライン確度。初期設定値は10万分の4です。
--cnv-slm-fw	放出されるCNVの最小データポイント数。初期設定は0ですが、これは1つの設計プローブを持つセグメントが事実上放出できることを意味します。
--cnv-slm-omega	実験的および生物学的分散間の相対重量を調整するスケーリングパラメーター。初期設定は0.3です。
--cnv-slm-stepeta	距離ノーマライゼーションパラメーター。初期設定は10000です。このオプションは、HSLMのみで有効です。

セグメンテーションメソッドに関係なく、セントロメアのような、深度データが使用できない大きいギャップにわたって、初期セグメントが分割されます。

[2] Orlandini V, Provenzano A, Giglio S, Magi A. SLMSuite: a suite of algorithms for segmenting genomic profiles. BMC Bioinformatics. 2017;18(1). doi:10.1186/s12859-017-1734-5

ターゲットセグメンテーション(セグメントBED)

ターゲットパネルのアプリケーションでは、--cnv-segmentation-bedを指定することにより、間隔で実行されるセグメンテーションおよびコールを制限できます。例えば、指定した間隔は、ターゲットアッセイに一致した遺伝子境界に対応する場合があります。このセグメンテーションモードは正常サンプルのパネルでのみサポートされており、--cnv-target-bedを付随する必要があります。正常サンプルのパネルの生成ステップ時に--cnv-segmentation-bedも指定することにより、解析時のすべての間隔境界を一致させます。正常サンプルのパネルの生成の詳細については、143 ページの 「正常サンプルのパネル」 を参照してください。

BEDファイルの推奨されるフォーマットには、4つの列と1つのヘッダーが含まれています。4つの列は、contig、start、stop、およびnameです。name列は遺伝子の名前を表しており、BEDファイル内でユニークにする必要があります。nameは出力VCFで使用され、INFO/SEGIDフィールドでセグメント識別子としてアノテーションされます。以下は、推奨されるフォーマットのファイルの例です：

```
contig start stop name
chr1 40356094 40372764 MYCL1
chr1 115245083 115261621 NRAS
chr1 204485504 204526342 MDM4
chr2 16075981 16090656 MYCN
chr2 29416087 30143527 ALK
chr3 12626010 12704516 RAF1
chr3 138374228 138478187 PIK3CB
chr3 178866307 178952154 PIK3CA
chr3 195776751 195806640 TFRC
```

3列BEDファイルを使用する場合は、ヘッダーおよびnameフィールド値を含めません。3列BEDファイルには、contig、start、およびstop値のみを含める必要があります。この場合、セグメント識別子は座標フィールドから自動生成されます。

クオリティスコアリング

クオリティスコアは、ヘヴィーテイル確度分布（整数のコピー数ごとに1つ）とイベント長の重み付けを組み合わせて使用する確率論的モデルを使用して計算されます。ノイズ分散を推定します。出力VCFには、コールされた増幅（二倍体座位の場合CN > 2）、欠失（二倍体座位の場合CN < 2）、またはコピーニュートラル（二倍体座位の場合CN = 2）イベントで信頼度を測定するPhred値のメトリクスが含まれています。

スコアリングアルゴリズムでは、DeNovo CNV検出パイプラインへの入力である厳密なコピー数クオリティスコアも計算します。

除外BEDフィルタリング

CNVコーラーに除外BEDを入力して、解析から領域をフィルタリングして除外できます。ライブラリー調製、シーケンス、またはマッピング問題が原因で、ゲノムで問題があるとして知られている特定の領域が存在する場合、除外BEDの入力は有効です。下流の解析に役に立つように、一般的なCNVを指定する大きい間隔を除外することもできます。`cnv-exclude-bed`を使用して、除外BEDファイルを作成できます。DRAGENでは、除外BEDを提供していません。除外する間隔を、標準の3列BEDフォーマットでフォーマットする必要があります。

除外BEDの間隔を、元のターゲットカウント間隔と比較します。重複が`cnv-exclude-min-overlap`より大きい場合、解析からターゲットカウント間隔が除外されます。`*.target.counts.gz`ファイルには依然として間隔が含まれているため、元のリードカウントを検査できます。ノーマライゼーションステージでは間隔を除去します。`*.tn.tsv.gz`ファイルでは、除去した間隔を除外します。

除外した間隔では、CNVコールが間隔に広がっていないことを保証していません。領域の側面に位置する十分なデータが存在する場合、セグメンテーションステージとの結合により、除外した間隔にわたるコールを生成する場合があります。ただし、コールでは、除外した間隔からのリードカウントを考慮しません。除外した間隔の説明を、`*.excluded_intervals.bed.gz`ファイルに表示できます。

出力ファイル

DRAGENホストソフトウェアでは、多数の中間ファイルを生成します。`*.seg.called.merged`は、増幅および欠失イベントを含む最終のコールファイルです。

セグメントファイルに加えて、DRAGENでは標準のVCFフォーマットのコールを放出します。初期設定では、VCFファイルには、コピー数増加および減少イベントのみが含まれています。コピーニュートラルセグメントでは、`*.seg.called.merged`ファイルを参照します。出力VCFにコピーニュートラル（REF）コールを含めるには、*--cnv-enable-ref-calls*をtrueに設定します。

`*.seg.called.merged`ファイル、およびデバッグと解析に役に立つ出力ファイルの使用方法の詳細については、134 ページの 「シグナルフロー解析」 を参照してください。

CNV VCFファイル

CNV VCFファイルは、標準のVCFフォーマットに従っています。CNVイベントの表示方法対構造多型の表示方法の性質により、すべてのフィールドが適用できるとは限りません。一般に、イベントに関して多数の情報が使用できる場合、情報はアノテーションされます。DRAGEN CNV VCFの一部のフィールドは、CNVに特有です。

以下は、CNVに特有のヘッダー行の例です。

```
##fileformat=VCFv4.2
##CoverageUniformity=0.402517
##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
##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"> ##INFO=<ID=REFLEN,Number=1,Type=Integer,Description="Number
of REF positions included in this record">
##INFO=<ID=SVLEN,Number=.,Type=Integer,Description="Difference in length
between REF and ALT alleles">
##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=GT,Number=1,Type=String,Description="Genotype">
##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">
```

POS列は、バリアントの開始位置です。VCF仕様に従って、ALTアリルのいずれかが<DEL>のような記号アリルの場合、パディング塩基が必須であり、POSは多型に先立つ塩基の座標を示しています。VCFのすべての座標は1-basedです。

ID列は、イベントを表すのに使用されます。IDフィールドは、イベントタイプとイベントの座標をコード化します。

REF列には、すべてのCNVイベントのNが含まれています。

ALT列では、CNVイベントのタイプを指定します。VCFに含まれているのはCNVイベントのみであるため、DELまたはDUPエントリーが使用されます。

QUAL列には、ハードフィルタリングで使用される、CNVイベントの推定クオリティスコアが含まれています。

CNVイベントがすべてのフィルターに合格した場合、FILTER列にはPASSが含まれており、合格していない場合、列には不合格のフィルターの名前が含まれています。

INFO列には、イベントを表す情報が含まれています。REFLENエントリーは、イベントの長さを示しています。SVTYPEエントリーは常にCNVです。ENDエントリーは、イベントの終了位置を示しています。セグメントBEDファイルを使用する場合、セグメント識別子が入力からSEGIDフィールドに引き継がれます。

FORMATフィールドは、ヘッダーに記述されています。

- GT：遺伝型
- SM：セグメント平均値の線形コピー割合
- CN：推定コピー数
- BC：領域のビン数
- PE：開始および終了ブレークポイントで不適当なペアエンドリードの数

生殖細胞系列コピー数コールでは、各ハプロタイプのコピー数ではなく全体のコピー数を決定するため、CNが2以上の際に遺伝型タイプフィールドには、二倍体領域の欠落値が含まれています。以下は、さまざまなVCFエントリーのGTフィールドの例です。

二倍体か一倍体か	ALT	FORMAT:CN	FORMAT:GT
二倍体	.	2	./.
二倍体	<DUP>	> 2	./1
二倍体	<DEL>	1	0/1
二倍体	<DEL>	0	1/1
一倍体	.	1	0
一倍体	<DUP>	> 1	1
一倍体	<DEL>	0	1

カバレッジ均一性

DRAGEN CNVパイプラインでは、サンプルデータのクオリティの基準を提供しています。WGSセルフノーマライゼーションメソッドを使用する場合、VCFヘッダーには追加のCoverageUniformityメトリクスが存在しています。メトリクスは、生殖細胞系列サンプルでのみ使用できます。CNVパイプラインでは、ノーマライゼーション後のターゲットカウントは、個別にまったく同様に分布している（IID）と仮定しています。最もハイクオリティのWGSサンプルのカバレッジは、CNVコーラーが精確なコールを生成するのに十分均一ですが、一部のサンプルはIID仮定に違反しています。ライブラリー調製またはサンプルコンタミネーション時の問題により、いくつかの極端な外れ値やターゲットカウントのうねりが発生する場合があり、これにより偽陽性CNVコールが大量に生成されることがあります。CoverageUniformityメトリクスは、サンプルのローカルカバレッジ相関の程度を定量して、クオリティが低いサンプルの同定を支援します。

このメトリクスの値が大きいということは、サンプルのカバレッジの均一性が低いという意味であり、これはサンプルにランダムではないノイズが含まれていることを示しており、クオリティが低いとみなされる場合があります。CoverageUniformityメトリクスは、*cnv-interval-width*設定やサンプル平均カバレッジのようなサンプルクオリティではなく、要因によって決まります。DRAGENでは、このスコアを使用して、類似した平均カバレッジおよび同じコマンドラインオプションからのサンプルのクオリティを比較することを推奨しています。このため、DRAGEN CNVではメトリクスのみを提供しており、それに基づいた措置は講じません。

CNVメトリクスファイル

DRAGEN CNVは、CSVフォーマットでメトリクスを出力します。出力は、DRAGENでレポートされるQCメトリクスの一般的な規則に従っています。CNVメトリクスは、ファイル拡張子が`*.cnv_metrics.csv`であるファイルへの出力です。以下のリストに、CNVランからの出力であるメトリクスを示します。

性ジェノタイパーメトリクス

- サンプルの推定した性核型で、信頼度メトリクスは、0.0～1.0の範囲です。サンプルの性が指定されている場合、このメトリクスは0.0です。
- 正常サンプルのパネルを使用する場合、すべてのパネルサンプルもレポートされます。

CNVサマリーメトリクス

- 使用しているリファレンスゲノムの塩基。
- ゲノムにわたる平均アライメントカバレッジ。
- 処理されるアライメント記録数。
- フィルタリングされた記録数（合計）。
- フィルタリングされた記録数（重複による）。
- フィルタリングされた記録数（MAPQによる）。
- フィルタリングされた記録数（マップされていないことによる）。
- ターゲット間隔数。
- 正常サンプル数。
- セグメント数。
- 増幅数。
- 欠失数。
- PASS増幅数。
- PASS欠失数。

視覚化およびBigWigファイル

既知の真のセットで解析を実行する際には、パイプラインステージからの中間出力ファイルを使用できます。これらのファイルを解析して、微調整にオプションに役立てることができます。

すべてのファイルの構造は、オプションのヘッダー行があるBEDファイルに類似しています。

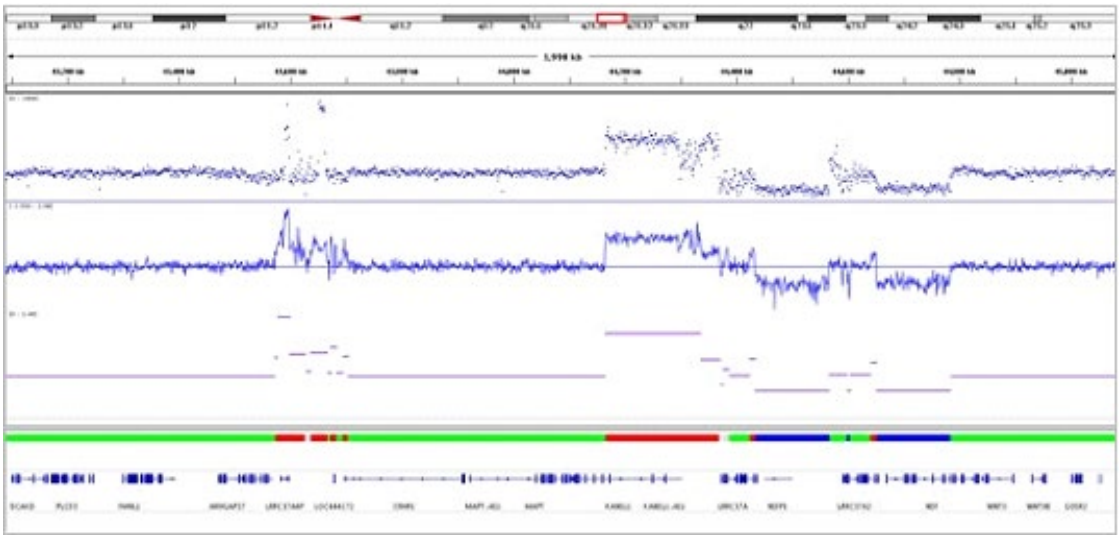
オプション	説明
**.target.counts.gz*	ターゲット間隔ごとのリードカウント数が含まれています。これは、BAMまたはCRAMファイルのアライメントから抽出した処理前のシグナルです。フォーマットは、ケースサンプルと正常サンプルのいずれかのパネルの両方で同一です。`target.counts.diploid`ファイルのbigWig表現も存在しており、これは処理前カウントではなく正常ploidyレベル2にノーマライズされています。
**.tn.tsv.gz*	ターゲット間隔ごとのケースサンプルのタンジェントノーマライズシグナル。このファイルには、対数ノーマライズしたコピー割合シグナルが含まれています。0.0からの強いシグナル偏差はCNVイベントの可能性を示しています。
**.seg.called.merged*	セグメンテーションアルゴリズムから生成されたセグメントが含まれています。
**.cnv.vcf.gz*	イベントを示す出力CNV VCFファイル。

追加の同等bigWigおよびgffファイルを生成するには、*--enable-cnv-tracks*オプションをtrueに設定します。これらのファイルは、RefSeq遺伝子のような、その他のトラックとともにIGVにロードできます。これらのトラックと公に使用できるトラックをともに使用することにより、コールの解釈が簡単になります。トラックがDRAGEN CNVで生成されている場合、DRAGENはIGVセッションXMLファイルを自動生成します。`*.cnv.igv_ session.xml`は、解析のためにIGVに直接ロードできます。

以下のIGVトラックは、出力IGVセッションファイルで自動的に入力されます：

オプション	説明
.target.counts.bw	ターゲットカウントビンのBigWig表現。IGVのトラックビューを棒グラフまたはポイントに設定することを推奨します。
**.improper_pairs.bw*	不適当なペアカウントのBigWig表現。IGVのトラックビューを棒グラフに設定することを推奨します。
**.tn.bw*	タンジェントノーマライズシグナルのBigWig表現。IGVのトラックビューをポイントに設定することを推奨します。
**.seg.bw*	セグメントのBigWig表現。IGVのトラックビューをポイントに設定することを推奨します。
**.cnv.gff3*	CNVイベントのGFF3表現。DELイベントは青色として、DUPイベントは赤色として表示されます。フィルタリングされたイベントは薄灰色です。イベントを選択するとウィンドウが立ち上がり、アノテーションの詳細が表示されます。

図 8　IGVの例



IGVセッションXML

IGVセッションXMLファイルには、DRAGENで生成されたトラックファイルが事前に入力されています。セッションファイルは、コマンドラインで指定した`--ref-dir`の名前を比較することにより、IGVインストールの標準リファレンスゲノムに最もよく一致するリファレンスゲノムをロードします。標準のUCSCヒトリファレンスゲノムは自動検出されますが、標準のリファレンスゲノムからの変動は自動検出されない場合があります。ゲノム検出を編集するには、IGVにロードする前に解析対象とするリファレンスゲノムへの`Session`要素の`genome`属性を変更します。IGVで使用されるリファレンス識別子は、ゲノムの実際の名前とは異なる場合があります。以下は、編集するセッションファイルの例です。

```
<?xml version="1.0" encoding="utf-8"?>
<Session genome="b37" hasGeneTrack="false" hasSequenceTrack="true"
version="8">
<Resources>
<Resource path="example.cnv.gff3"/>
<Resource path="example.cnv.excluded_intervals.bed.gz"/>
<Resource path="example.target.counts.bw"/>
<Resource path="example.improper.pairs.bw"/>
<Resource path="example.tn.bw"/>
<Resource path="example.seg.bw"/>
</Resources>
<Panel height="500" width="1200" name="DataPanel">
...
</Panel>
</Session>
```

間隔ファイルの除外

精度を改善するために、DRAGEN CNVパイプラインでは、1つ以上のターゲット間隔が少なくとも1つのクオリティ要件で不合格であった場合、ゲノム間隔を除外します。除外した間隔は、`*.excluded_intervals.bed.gz`ファイルにレポートされます。ファイルは、CNV解析でコール可能ではないゲノムの領域を同定し、間隔が4番目の列で除外された理由を記述します。以下は、除外で可能性のある理由です。

除外理由	説明	コマンドライン オプション
NON_KMER_UNIQUE	ユニークではないKmer塩基は、間隔の50%より大きくなります。	該当なし。この理由は、セルフノーマライゼーションモードにのみ適用されます。
EXCLUDE_BED	間隔は、閾値より大きい除外BEDと重複しています。	`--cnv-exclude-min-overlap`
PON_MAX_PERCENT_ZERO_SAMPLES	カバレッジが0のPONサンプル数が閾値より大きくなっています。	`--cnv-max-percent-zero-samples`

除外理由	説明	コマンドライン オプション
PON_TARGET_FACTOR_THRESHOLD	間隔の中央値カバレッジが、全体の中央値カバレッジの閾値より低くなっています。	`--cnv-target-factor-threshold`
PON_MISSING_INTERVAL	PONでターゲット間隔が見つかりません。	該当なし。

出力およびフィルタリングオプション

出力およびフィルタリングオプションにより、CNV出力ファイルを制御します。

オプション	説明
--cnv-exclude-bed	CNV解析から除外する間隔を示すBEDファイルを指定します。ターゲット間隔が、cnv-exclude-min-overlapを超える除外BEDファイルから指定された領域と重複している場合、ターゲット間隔は抑制されます。
--cnv-exclude-min-overlap	ターゲット間隔と除外した領域間の重複量の閾値をフィルタリングするための断片を指定します(0.5)。
--cnv-enable-plots	CNVパイプラインの一部としてプロットを生成します。初期設定はfalseです。高解像度間隔(1000 bp未満)でWGS CNV解析を実行する場合、プロットの生成が完了するのに長い時間を要する場合があります。Illuminaでは、初期設定(無効)を使用することを推奨しています。
--cnv-enable-ref-calls	出力VCFファイルのコピーニュートラル(REF)コールを放出します。単一WGS CNV解析での初期設定はtrueです。
--cnv-enable-tracks	表示するためにIGVにインポートできるトラックファイルを生成します。このオプションが有効な場合、出力バリアントコールの`*.gff`ファイル、およびタンジェントノーマライズシグナルの`*.bw`ファイルが生成されます。初期設定はtrueです。

オプション	説明
--cnv-filter-bin-support-ratio	全体のイベント長に対して、ビンをサポートする範囲が指定した割合よりも小さい場合、候補イベントをフィルタリングして除外します。初期設定の割合は0.2です(20%サポート)。例として、イベントがコールされて長さが100000 bpであるが、コールをサポートするターゲット間隔ビンがわずか合計15000 bp(15000/100000 = 0.15)の範囲の場合、間隔はフィルタリングで除外されます。
--cnv-filter-copy-ratio	レポートされるイベントが出力VCFファイルでPASSとマーキングされている、約1.0を中心とする最小コピー割合閾値を指定します。初期設定値は0.2ですが、これはCR = 0.8より小さいかまたはCR = 1.2より大きいコールを引き起こします。
--cnv-filter-length	レポートされるイベントが出力VCFファイルでPASSとマーキングされている、塩基の最小イベント長を指定します。初期設定は10000です。
--cnv-filter-qual	レポートされるイベントが出力VCFファイルでPASSとマーキングされている、QUAL値を指定します。独自のアプリケーションデータに従って、パラメーター値を調整する必要があります。
--cnv-min-qual	レポートされる最小QUALを指定します。初期設定は3です。
--cnv-max-qual	レポートされる最大QUALを指定します。初期設定は200です。
--cnv-ploidy	正常のploidy値を指定します。このオプションは、出力VCFファイルに放出されるコピー数値の推定にのみ使用されます。初期設定は2です。
--cnv-qual-length-scale	バイアス重み付け係数を指定して、長いセグメントのQUAL推定値を調整します。これは高度なオプションであり、変更する必要はありません。初期設定は0.9303(2-0.1)です。
--cnv-qual-noise-scale	バイアス重み付け係数を指定し、サンプル変動に基づいてQUAL推定値を調整します。これは高度なオプションであり、変更する必要はありません。初期設定は1.0です。

同時CNVおよびスモールバリアントコール

DRAGENは、FASTQサンプルのマップとアライメントを実行して、データを直接下流のコーラーにストリームできます。入力がFASTQサンプルの場合、単一のサンプルはCNVとスモールVCの両方を通して実行できます。初期設定では、これによりセルフノーマライゼーションを引き起こします。

通常のDRAGENマップ/アライメントワークフローを通してFASTQサンプルを実行してから、引数を追加してCNV、VC、または両方を有効にします。スタンドアロンワークフローでCNVに適用されるオプションは、ここでも適用できます。

以下の例は、さまざまなコマンドを示しています。

CNVによるFASTQのマップ/アライメント

```
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
```

VCによるFASTQのマップ/アライメント

```
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
```

CNVおよびVCによるFASTQのマップ/アライメント

```
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
```

CNVおよびVCへのBAM入力

```
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
```

サンプル相関と性ジェノタイパー

ターゲットカウントステージまたはノーマライゼーションステージを実行する際には、DRAGEN CNVパイプラインは、ランのサンプルに関する以下の情報も提供します。

- ケースサンプルと正常サンプルのいずれかのパネル間のリードカウントプロファイルの相関メトリクス。信頼性の高い解析では0.90より大きい相関メトリクスを推奨しますが、ソフトウェアにより実施される厳しい制限はありません。
- ランの各サンプルの予測される性。性は、性染色体および常染色体のリードカウント情報に基づいて予測されます。カウントの中央値は、常染色体、X染色体、およびY染色体の画面に出力されます。

パイプラインの実行時に、結果が画面に出力されます。例えば：

```
==========================================
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
```

使用しているサンプルの予測される性も、`*.cnv_metrics.csv`出力ファイルに出力されます。

正常サンプルのパネルでは、予測される性を使用して、性染色体のノーマライゼーションに利用されるパネルサンプルを決定します。サンプルの推定される性がUNDETERMINEDの場合、サンプルの性はFEMALEに設定されます。

*–sample-sex*オプションで、サンプルの予測される性をオーバーライドできます。

マルチサンプルCNVコール

--cnv-inputオプション（サンプルごとに1つ）で指定されたタンジェントノーマライズカウントファイル（*.tn.tsv.gz）からマルチサンプルCNVコールを開始できます。マルチサンプルCNV解析は、ジョイントセグメンテーションを使用することで、コピー数可変セグメントの検出感度を増加させます。同定された各コピー数可変セグメントに対して、各サンプルのコピー数遺伝型は、単一のVCFエントリー中に出力され、アノテーションおよび解釈が容易になります。

WGSおよびWESワークフローでは、マルチサンプルCNVコールがサポートされています。

トリオ解析を実行するコマンドラインの例を次に示します。

```
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>
```

De Novo CNVコールオプション

すべてのインプットサンプルが同じ単一サンプルワークフローを介して、同じ間隔であることを確認します。サンプルがWESインプットの場合、同じ正常サンプルのパネルを使用してサンプルを生成する必要があり、全サンプルの常染色体の間隔が一致している必要があります。

DeNovo CNVコールでは、次のオプションが使用されます：

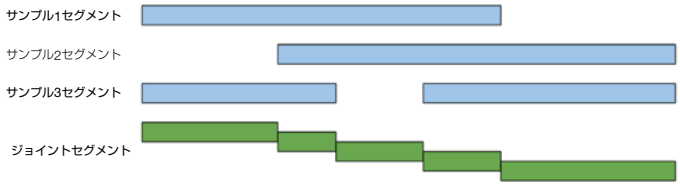
オプション	説明
--cnv-input	DeNovo CNVコールの場合、単一サンプル実行からのインプットするタンジェントノーマライズシグナルファイル（*.tn.tsv）を指定します。このオプションは、インプットサンプルごとに1回ずつ、複数回指定できます。
--cnv-input	DeNovo CNVコールの場合、単一サンプル実行からのインプットするタンジェントノーマライズシグナルファイル（*.tn.tsv）を指定します。このオプションは、インプットサンプルごとに1回ずつ、複数回指定できます。

オプション	説明
--cnv-filter-de-novo-qual	発端者サンプル中の推定イベントがDeNovoとしてマークされるPhred値の閾値。初期設定値は0.125です。
--pedigree-file	インプットサンプル間の関係を指定するpedigreeファイル。

ジョイントセグメンテーション

まず、CNVコールを各サンプルに対して独立に実行します。次いで、ジョイントセグメンテーションでは、各単一サンプル解析からのコピー数可変セグメントを使用して、ジョイントコピー数可変セグメント一式を導出します。このジョイントセグメント一式は、全サンプルのコピー数可変セグメントからすべてのブレークポイントの和を取ることによって簡単に決定されます。この結果、異なるサンプル間で部分的に重複するセグメントが分割されます。

図 9　重複セグメント



ジョイントセグメンテーションに続いて、ジョイントセグメントを使用して、各サンプルに対して独立してコピー数コールが再度実行されます。セグメントは単一サンプル解析と同様にマージすることができますが、各ジョイントセグメントは単一エントリーとしてマルチサンプルVCFに放出されます。サンプルのマージされたセグメントからのクオリティスコア（VCF内のQS）は、該当する場合、コールのフィルタリングに使用されます。サンプルコールは、マルチサンプルVCF内のサンプルのFTフィールドを使用してフィルタリングされます。マルチサンプルVCFのQUAL列は常に欠落しています（すなわち「.」）。マルチサンプルVCFのFILTER列は、サンプルのFTフィールドがいずれも「PASS」でなければ「SampleFT」であり、サンプルのFTフィールドのいずれかが「PASS」であれば「PASS」です。

De Novoコールステージ

de novoイベントとは、発端者のゲノムの特定の座位に遺伝型が存在し、その遺伝型が両親からの標準的なメンデル遺伝の結果ではない場合と定義されます。de novoコールステージでは、マルチサンプル解析の各トリオの発端者において推定de novoイベントを同定します。場合によっては、これらの推定されるde novoイベントは現実のものであるかもしれませんが、それらはシーケンシングや解析のアーティファクトから生じることもあります。その結果、推定される各de novoイベントにde novoのクオリティスコアが割り当てられ、クオリティの低いde novoイベントをフィルタリングして除外するために用いられます。トリオを指定するには、.pedファイルを*--pedigree-file*オプションで指定します。複数のトリオを指定でき（例：クワッド解析）、有効なトリオがすべて処理されます。

トリオのそれぞれのジョイントセグメントについて、de novoコーラーは、そのコピー数の遺伝型についてメンデル遺伝の不一致があるかどうかを決定します。CNVコーラーは、与えられた二倍体セグメントの各アリルのコピー数を同定しません。これは、親遺伝型のアリル組成の可能性について仮定することを意味します。

割り当てられたコピー数が2以上の場合、親ゲノムの二倍体領域にはコピー数0のアリルは存在しない（性別に依存する）と仮定します。これにより、次のように簡略化されます：

親のコピー数の遺伝型	可能なコピー数のアリル	推定される可能なコピー数のアリル
2	0/2、1/1	1/1
3	0/3、1/2	1/2
4	0/4、1/3、2/2	1/3、2/2
N	x/(N-x) for x <= N/2	x/(N-x) for 1 <= x <= N/2

以下は、これらの仮定を用いた二倍体領域のコピー数が一致する遺伝型と一致しない遺伝型の例です：

母親のコピー数	父親のコピー数	発端者のコピー数	メンデルの法則の一致?
2	2	2	はい
2	2	1	いいえ
3	2	4	いいえ
3	2	2	はい
2	0	2	いいえ

ジョイントセグメントにメンデル遺伝の不一致がある場合、Phred値のde novoのクオリティスコア（VCFのDQフィールド）は、トリオの各サンプルのコピー数状態（「クオリティスコアリング」のセクションを参照）それぞれの尤度と、トリオの遺伝型の事前確率を用いて計算されます：

```
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)))
```

ここで：

- CNm =母親のコピー数
- CNf =父親のコピー数
- CNp =発端者のコピー数
- p(CNm,CnF,CNp) =トリオ遺伝型の事前確率

VCFのDNフィールドは、各セグメントのde novoステータスを示すために使用されます。設定可能な値は次のとおりです：

- Inherited：コールされるトリオ遺伝型はメンデル遺伝と一致しています。

- LowDQ：コールされるトリオ遺伝型はメンデル遺伝と一致しておらず、DQはde novoクオリティ閾値未満です（初期設定値は0.125）。
- DeNovo：コールされるトリオ遺伝型はメンデル遺伝と一致しておらず、DQはde novoクオリティ閾値以上です（初期設定値は0.125）。

マルチサンプルCNV VCF出力

マルチサンプルCNV VCFにおけるレコードは、単一サンプルの場合とは若干異なります。主な違いは次のとおりです：

- レコードごとのエントリーは、すべてのインプットサンプルのブレークポイントの和集合内のセグメントに分割されます。これは、VCF全体により多くのエントリーがあることを意味します。
- QUAL列は使用されず、値は「.」です。サンプルごとのクオリティは、QSタグとともにSAMPLE列に引き継がれます。
- 個々のSAMPLE列のいずれかがPASSの場合、FILTER列はPASSを示します。それ以外の場合は、SampleFTを示します。
- サンプルごとのアノテーションは、元のコールから引き継がれます。単一サンプルフィルターはサンプルレベルで適用され、FTアノテーションで放出されます。

さらに、有効なpedigreeが使用される場合、de novoコールが実行され、以下の2つのアノテーションが発端者のサンプルに追加されます。

```
##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 > 0.125000">
```

VCFには多くのエントリーが含まれますが、ジョイントセグメンテーションステージのために、de novoイベントの数は、DNおよびDQのアノテーションを有するエントリーの抽出によって見つけることができます。これらのレコードも抽出され、de novoコールの場合はGFF3に変換されます。

体細胞CNVコール

ワークフローに応じて、次の体細胞CNVコールを使用できます。

- 全ゲノムシーケンス（WGS）には、体細胞WGS CNVコーラーを使用します。詳細については、167 ページの 「体細胞WGS CNVコール」 を参照してください。
- 全エクソームシーケンス（WES）には、体細胞WES CNVコーラーを使用します。詳細については、164 ページの 「体細胞WES CNVコール」 を参照してください。

体細胞WES CNVコール

体細胞全エクソームシーケンス（WES）および体細胞ターゲットパネルでは、正常サンプルのパネルをリファレンスベースラインとして使用すれば、コピー数バリアントに関する知見を得ることができます。レポートされるイベントは、ノーマライズされたコピー割合値および期待されるリファレンスベースラインレベルからの偏差のみに基づきます。このワークフローは、ターゲット遺伝子の増減の検出のみを必要とするアプリケーションに有用です。体細胞WES CNVモデルは生殖細胞系列WES CNVモデルと類似していますが、異なるクオリティスコアリングおよびコールモデルを利用します。

次のいずれかのインプットオプションを使用します。

- `--tumor-fastq1`および`--tumor-fastq2`：FASTQファイルを指定します
- `--tumor-bam-input`：既存のBAMファイルを指定します
- `--tumor-cram-input`：既存のCRAMファイルを指定します

体細胞WES CNVコーラーには、正常サンプルのパネルが必要です。正常サンプルのパネルは、適切なノーマライゼーションを可能にする上流プロセスの本質的なバイアスを測定するのに役立ちます。正常サンプルのパネルを生成する方法については、143 ページの 「正常サンプルのパネル」 を参照してください。正常サンプルのパネルは、解析中のケースサンプルとよくマッチするはずです。

マッチする正常サンプルが利用可能な場合、サンプルを正常サンプルのパネルに含めることができます。マッチする正常サンプルが使用可能かどうかにかかわらず、ワークフローは変更されません。

コマンドラインの例

次のコマンドラインの例では、WESデータに対して体細胞解析を実行します。

```
dragen \
-r <HASHTABLE> \
--output-directory <OUTPUT> \
--output-file-prefix <SAMPLE> \
--enable-map-align false \
--enable-cnv true \
--tumor-bam-input <TUMOR_BAM> \
--cnv-normals-list <NORMALS> \
--cnv-target-bed <BED> \
```

キャプチャーキットの標的遺伝子で体細胞ターゲットパネルを使用する場合は、`cnv-segmentation-bed`を指定し、`cnv-segmentation-mode=bed`を使用することでセグメンテーションをバイパスできます。このオプションを使用すると、セグメンテーションBED内の全イベントが出力VCFにレポートされます。セグメンテーションBEDファイルに関する詳細については、149 ページの 「ターゲットセグメンテーション（セグメントBED）」 を参照してください。

次のコマンドラインの例では、ターゲットパネルで体細胞解析を実行します。

```
dragen \
-r <HASHTABLE> \
--output-directory <OUTPUT> \
--output-file-prefix <SAMPLE> \
--enable-map-align false \
```

```
--enable-cnv true \
--tumor-bam-input <TUMOR_BAM> \
--cnv-normals-list <NORMALS> \
--cnv-target-bed <BED> \
--cnv-segmentation-bed <SEGMENT_BED> \
--cnv-segmentation-mode bed \
```

クオリティスコアリングとコール

体細胞WES CNVコーラーは、ケースサンプルのノーマライズコピー割合と正常サンプルのパネルとの間の2サンプルt検定を用いてクオリティスコアを計算します。コーラーはセグメントごとにp値を計算します。次に、p値をPhred値のスコアに変換します。クオリティスコアが`cnv-filter-qual`値より高い場合、イベントはDUP/DELとして記録されます。初期設定値は90です。コピーニュートラルイベントの場合、コーラーはクオリティスコアを`1-p`として計算します。

出力VCFには、`QUAL`フィールドにクオリティスコアが含まれます。

体細胞WES CNV出力

体細胞WES CNV VCFファイルは、標準VCFフォーマットに従っており、生殖細胞系列CNV VCF出力とは以下の相違点があります。生殖細胞系列CNV VCF出力について詳しくは、150 ページの 「CNV VCFファイル」 を参照してください。

`FILTER`フィールドは、CNVが、特定された検出下限（LOD）閾値よりもニュートラルPONカウントに近いかどうかを決定するための追加フィルターを含みます。`LoDFail`フィルターは次のように計算されます。ここで、LoDは`cnv-filter-limit-of-detection`を使用して設定されます（初期設定値：0.2）。

$$PON \times (1 \pm LoD)$$

`FORMAT`フィールドはヘッダーセクションで説明されており、体細胞WES CNV VCFに`CN`エントリーが含まれていないことを除き、生殖細胞系列CNV VCF出力と同一です。体細胞WES CNVコーラーは腫瘍の純度分画を推定せず、コピー数の推定もできません。詳細については、167 ページの 「腫瘍の純度と倍率変化」 を参照してください。

以下は、セグメンテーションBEDが体細胞WES CNVコールに含まれたVCF出力の例です。

```
chr7 92243233 DRAGEN:REF:chr7:92243233-92462639 N . 3 PASS
END=92462639;REFLEN=219407;SEGID=CDK6 GT:SM:BC:PE ./.:1.0257:28:125,0
chr7 116339136 DRAGEN:GAIN:chr7:116339137-116436180 N <DUP> 200 PASS
SVLEN=97044;SVTYPE=CNV;END=116436180;REFLEN=97044;SEGID=MET GT:SM:BC:PE
./1:1.37061:36:2287,989
chr7 140434395 DRAGEN:REF:chr7:140434395-140624505 N . 3 PASS
END=140624505;REFLEN=190111;SEGID=BRAF GT:SM:BC:PE ./.:0.977961:38:73,0
```

腫瘍の純度と倍率変化

WESおよびターゲットパネルがまばらであるため、b-アリルデータは腫瘍の純度の正確な推定には不十分です。Somatic WES CNVコーラーは、倍率変化としても知られるコピー割合のみをレポートします。倍率変化は、セグメント平均の線形コピー割合として`FORMAT/SM`フィールドにコード化されます。

腫瘍の純度が既知の場合には、次の計算を用いて、レポートされた倍率変化からサンプル中の遺伝子またはセグメントのploidyを推測することができます。

$$\text{コピー数} = \frac{[(200^{\text{倍の倍率変化}})-(2\times[100-\text{腫瘍の純度\%}])]}{\text{腫瘍の純度\%}}$$

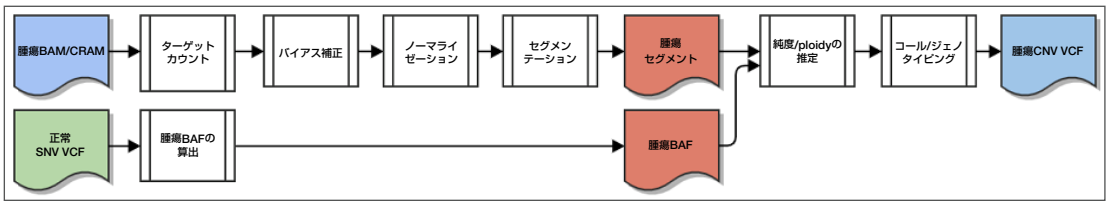
例えば、腫瘍の純度がMETについて30%であり、2.2xの倍率変化を伴う場合、サンプル中にはMET DNAが10コピー存在します。

体細胞WGS CNVコール

体細胞コピー数異常およびヘテロ接合性の消失を伴う領域を検出するために、生殖細胞系列SNVを含むVCFを有する腫瘍サンプルに対してDRAGEN CNV Callerを実施します。出力ファイルはVCFファイルです。生殖細胞系列CNVコーラーのコンポーネントは、腫瘍の純度およびploidyを推定する体細胞モデリングのコンポーネントを追加して体細胞アルゴリズムで再利用されます。

生殖細胞系列SNVは腫瘍中のb-アリル割合を計算するために用いられ、腫瘍サンプルを必要とするアリル特異的コピー数コールを可能にします。可能であれば、マッチする正常サンプル由来のスモールバリアントVCFの使用が望ましい（Tumor-Normalモード）ですが、マッチする正常サンプルが利用できない場合は、集団SNPのカタログを使用することができます（Tumor-onlyモード）。

図 10　体細胞CNV全ゲノムシーケンス（WGS）コーラーワークフロー



マッチする正常なサンプルが入手可能な場合、サンプルは最初に生殖細胞系列スモールバリアントコーラーを用いて処理します。この場合、生殖細胞系列ヘテロ接合性のSNV部位のみをb-アリル割合の決定に用います。マッチする正常サンプルが入手できない場合、集団SNPのb-アリル割合は、マッチする正常なヘテロ接合性座位の場合と同様に計算されますが、未知の生殖細胞系列遺伝型のバリアントとして扱われます。可能な遺伝型の割り当ては、アリル特異的なコピー数を決定するために統計的に統合されます。

マッチする正常モードでは、個体の生殖細胞系列コピー数の変化を含むVCFをオプションでインプットすることができます。これにより、体細胞全ゲノムシーケンス（WGS）CNV VCFにおいて生殖細胞系列CNVが別個のセグメントとして出力され、生殖細胞系列コピー数でアノテーションが付けられ、その領域に特異的な体細胞コピー数の変化があるかどうかが明確になります。

体細胞WES CNVコールオプション

次の体細胞WGS CNVコールコマンドラインオプションを使用できます。

オプション	説明
• --tumor-fastq1, • --tumor-fastq2, • --tumor-bam-input • --tumor-cram-input	腫瘍入力ファイルを指定します。
--cnv-normal-b-allele-vcf	マッチする正常SNV VCFを指定します。b-アリル座位の指定について詳しくは、170 ページの「B-アリル座位の指定」を参照してください。
--cnv-population-b-allele-vcf	集団SNPカタログを指定します。b-アリル座位の指定について詳しくは、170 ページの「B-アリル座位の指定」を参照してください。
--cnv-use-somatic-vc-baf	SNVコーラーを有効にしてTumor-Normalモードで実行している場合は、このオプションを使用して生殖細胞系列のヘテロ接合部位を指定します。b-アリル座位の指定について詳しくは、170 ページの「B-アリル座位の指定」を参照してください。
--sample-sex	既知の場合は、サンプルの性別を明記してください。サンプルの性別が指定されていない場合、コーラーは腫瘍アライメントからサンプルの性別を推定しようとします。
--cnv-normal-cnv-vcf	マッチする正常サンプルからの生殖細胞系列CNVを特定します。詳細については、171 ページの「Germline-Awareモード」を参照してください。
--cnv-use-somatic-vc-vaf	体細胞SNVからのバリアントアリル頻度（VAF）を用いて、サンプルの腫瘍モデルを選択するのに役立てます。詳細については、172 ページの「VAF-awareモード」を参照してください。
--cnv-somatic-enable-het-calling	不均質なセグメントのHET-callingモードを有効にします。詳細については、172 ページの「HET-callingモード」を参照してください。

以下は、マッチする正常なSNV VCFでTumor-Normal体細胞WGS CNVコーラーを実行するためのコマンドラインの例です。

```
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>
```

マッチする正常サンプルが使用できない場合は、CNVコールを無効にするか、Tumor-onlyモードで実行する必要があります。ミスマッチした正常サンプルを使用してTumor-Normalモードで実行すると、予期しない結果になります。次のコマンドラインの例では、集団SNV VCFを使用してTumor-onlyの体細胞WGS CNVコールを実行します。

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

次のコマンドラインの例では、体細胞SNVコーラーと同時にTumor Normal 体細胞WGS CNVコールを実行します。これにより、コマンド`cnv-use-somatic-vc-baf true`を使用して、SNVコーラーからマッチする正常な生殖細胞系列ヘテロ接合性部位を直接使用できます。

```
dragen \
-r <HASHTABLE> \
--output-directory <OUTPUT> \
--output-file-prefix <SAMPLE> \
--enable-map-align false \
--enable-cnv true \
--tumor-bam-input <TUMOR_BAM> \
--bam-input <NORMAL_BAM>
--enable-variant-caller true \
--cnv-use-somatic-vc-baf true \
--sample-sex <SEX>
```

マッチする正常サンプルとDRAGEN Germline解析からの出力も利用できる場合は、追加の機能を有効にできます。マッチする正常なサンプルが利用可能な場合は、次のコマンドラインの例を使用して、Germline-awareモードとVAF-awareモードを有効にします。Germline-awareモードおよびVAF-awareモードの詳細については、171 ページの 「Somatic WGS CNV Callingモード」 および172 ページの 「VAF-awareモード」 を参照してください。

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

```
--tumor-bam-input <TUMOR_BAM> \
--bam-input <NORMAL_BAM>
--enable-variant-caller true \
--cnv-use-somatic-vc-baf true \
--cnv-normal-cnv-vcf <CNV_NORMAL_VCF> \
--sample-sex <SEX>
```

ターゲットカウントおよびB-アリルカウント

ターゲットカウントステージとその出力は生殖細胞系列CNVコールの場合と同じです。リードカウントを含むターゲット間隔は、*.target.countsファイルに出力されます。b-アリルのカウントはリードカウントステージと並行して行われ、その値は*.baf.bedgraph.gzファイルに出力されます。このファイルは、視覚化のためにDRAGENによって生成された他のBigWigファイルとともにIGVにロードすることができます。

B-アリル座位の指定

体細胞WGS CNVコーラーは、腫瘍サンプルのb-アリル数をカウントするためにヘテロ接合性SNP部位の供給源を必要とします。使用できるモードは次のとおりです。

オプション	説明
cnv-normal-b-allele-vcf	マッチする正常SNV VCFを指定します。マッチする正常サンプルおよびマッチする正常SNV VCFが入手可能な場合に用います。このオプションを使用するには、マッチする正常サンプルをDRAGENの生殖細胞系列ワークフローで実行する必要があります。
cnv-population-b-allele-vcf	集団SNP VCFを指定します。マッチする正常サンプルが入手できず、Tumor-onlyモードで解析を実施しなければならない場合に使用します。
cnv-use-somatic-vc-baf	`true`に設定すると、DRAGENの生殖細胞系列ワークフローがバイパスされます。腫瘍インプットおよびマッチする正常インプットが入手可能な場合に使用します。このオプションを使用するには、体細胞SNVコーラーを有効にします。

マッチする正常サンプルSNV VCFを指定するには、*--cnv-normal-b-allele-vcf*オプションを使用します。VCFファイルは、フィルターを適用したDRAGEN生殖細胞系列スモールバリアントコーラーを介して、マッチする正常サンプルの処理によって得られます。通常、このファイル名には*.hard-filtered.vcf.gz拡張子が付きます。正常なサンプルでヘテロ接合性であると判定され、PASSとマークされたレコードすべてを用いて、腫瘍サンプルのb-アリルカウントを測定します。同等のgVCFファイル（*.hard-filtered.gvcf.gz）を使用することもできますが、レコードの数が多いため処理時間が大幅に長くなります。レコードのほとんどはヘテロ接合性部位ではありません。

集団SNP VCFを指定するには、*--cnv-population-b-allele-vcf*オプションを使用します。集団SNP VCFを得るには、dbSNP、1000ゲノムプロジェクト、その他の大規模なコホート発見の取り組みなどから、集団バリエーションの適切なカタログを作成します。高頻度のSNPのみを含めます。例えば、マイナーアリル集団の頻度が10%以上のSNPを含めて、ランタイムの影響を制限し、アーティファクトを低減します。各レコードの`INFO`セクションに`AF=<alt frequency>`を追加して、ALTアリル頻度を指定します。追加の`INFO`フィールドが存在する場合もありますが、DRAGENは`AF`フィールドのみを解析して使用します。*--cnv-population-b-allele-vcf*で指定される部位は、腫瘍ゲノムが由来する生殖細胞系列ゲノムにおいてヘテロ接合性またはホモ接合性のいずれかの可能性があります。

以下は、有効な集団SNPレコードの例です：

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

DRAGENでは、b-アリルVCFからのレコードを解析する際に、以下の要件が考慮されます：

- 単純なSNV部位のみです。
- レコードは、FILTERフィールドでPASSとマークされている必要があります。
- VCFに同じCHROMCとPOSを持つレコードがある場合、DRAGENは最初に発生したレコードを使用します。

腫瘍サンプルおよびマッチする正常なインプットが入手可能な場合は、*--cnv-use-somatic-vc-baf true*を使用します。体細胞SNVコーラーを有効にする必要があります。このオプションを使用する場合、DRAGENはマッチする正常インプットから生殖細胞系列のヘテロ接合性部位を決定し、腫瘍サンプルのb-アリルカウントを測定します。この情報は体細胞WGS CNVコーラーに渡され、体細胞ワークフロー全体を簡略化します。

--cnv-use-somatic-vc-bafを有効にするには、次のコマンドラインオプションを入力します。

- --tumor-bam-input <TUMOR_BAM>：腫瘍インプットを指定します
- --bam-input <NORMAL_BAM>：マッチする正常サンプルインプットを指定します
- --enable-variant-caller true：体細胞SNVバリアントコーラーを有効にします
- --cnv-use-somatic-vc-baf true：VC BAFを有効にします

Somatic WGS CNV Callingモード

Germline-Awareモード

マッチする正常サンプルから生殖細胞系列CNVを指定するには、--cnv-normal-cnv-vcfを使用します。指定された場合、正常サンプル中でPASSとマークされたCNVレコードは、腫瘍サンプルのセグメンテーション中に使用され、確実な生殖細胞系列CNV境界が体細胞アウトプット中の境界でもあることを確認します。リファレンスPloidyと比較して生殖細胞系列コピー数が変化しているセグメントは、体細胞モデルの選択から除外されます。

体細胞コピー数のコールおよびスコアリングの際に、生殖細胞系列コピー数を用いて、腫瘍サンプルの正常な混入割合からの予想される深度への寄与を修正します。このプロセスは、生殖細胞系列CNVの領域における体細胞コピー数のより正確な割り当てにつながります。次にDRAGENは、体細胞WGS CNV VCFエントリーに、生殖細胞系列コピー数（NCN）および生殖細胞系列CNVを有するセグメントの生殖細胞系列と比較した体細胞コピー数の差（SCND）でアノテーションを付けます。

VAF-awareモード

腫瘍がマッチする正常なランにおいて、スモールバリアントコーラーとCNVコーラーの両方が有効である場合、体細胞SNVの結果は、腫瘍サンプルの推定純度およびploidyに影響を及ぼす可能性があります。通過する体細胞SNVからのアリル深度値によって捕捉される体細胞SNVバリアントアリル頻度（VAF）は、腫瘍の純度、体細胞SNV座位における総腫瘍コピー数、および体細胞アリルを保有する腫瘍コピー数の組み合わせを反映します。同様のアリル深度を有する体細胞SNVのクラスターは腫瘍モデルに情報を提供します。

腫瘍のコピー数バリエーションが限られている場合、および/または多くの血液腫瘍の場合のようにCNVのほとんどがサブクローナルである場合、VAFは不正確な腫瘍モデルや信頼度の低い腫瘍モデルの推定を防ぐのに役立ちます。推定された腫瘍モデルが不正確であったり信頼度が低いと、誤ったコールやフィルタリングされたコールにつながる可能性があります。VAF情報は、明らかなクローンCNVであってもゲノム重複の有無を判定するのに役立ちます。

VAF情報を利用するには、腫瘍やマッチする正常なリードおよびアライメントのインプットに対して、スモールバリアントコールを用いて体細胞WGS CNVコールを実行します。

例えば、次のコマンドラインを使用できます：

```
--enable-vc=true --enable-cnv=true --tumor-bam-input <Tumor_BAM> --bam-input
<Norman_BAM>
```

VAFベースのモデリングは初期設定で有効になっています。VAFベースのモデリングを無効にするには、`--cnv-use-somatic-vc-vaf`を`false`に設定します。

HET-callingモード

初期設定で、DRAGENは、異なるサブクローン間で不均質（HET）であると推定されるコピー数を持つセグメントに対してHET-callingモードを使用します。統計モデルに基づいて、セグメント内のBAF値の深度が、最も近い整数コピー数に対して期待される深度から遠すぎる場合、セグメントは不均質であると考えられます。

HET-callingをオフにするには、コマンドラインで`--cnv-somatic-enable-het-calling=false`を指定します。

セグメントが不均質であるとみなされる場合、セグメントのアウトプットは次のように変化します。

- セグメントのINFOフィールドにHETタグが追加されます。
- CN値およびMCN値の少なくとも1つは、非REF値として与えられます。具体的には、CNFおよびMCNFに最も近い整数値として与えられます。整数値がREFコールをもたらす場合、CN値とMCN値の少なくとも1つは、最も近い非REF値に調整されます。
- 選択したCNおよびMCNに対して、ID、ALT、およびGTフィールドが適切に設定されます。
- QUALスコアは、セグメントが少なくともサンプルの一部において非リファレンスコピー数をもつという信頼度を反映します。
- CNQ値およびMCNQ値は、割り当てられたCN値およびMCN値が全サンプルにおいて真であるという信頼度を反映しているため、CNQ値およびMCNQ値の少なくとも1つは通常、5未満です。

体細胞WGS CNVモデル

腫瘍の純度および二倍体カバレッジ（ploidy）の選択は、体細胞WGS CNVコーラーの重要なコンポーネントです。体細胞WGS CNVコーラーは、多くの候補モデルを評価するグリッドサーチ法を用いて、腫瘍サンプル中の全セグメントにわたって観察されたリードカウントおよびb-アリルカウントを適合させようとします。各候補について対数尤度スコアが生成されます。対数尤度スコアは、`*.cnv.pure.coverage.models.tsv`ファイルに出力されます。体細胞WGS CNVコーラーは、純度、対数尤度の最も高いカバレッジペアを選択し、次いで、代替モデルと比較した選択モデルの相対的尤度に基づいて、モデル信頼度のいくつかの尺度を計算します。

選択されたモデルの信頼度が低い場合、出力VCFはlowModelConfidence `FILTER`で全レコードをマークし、VCFヘッダー内の推定腫瘍純度を`NA`に設定します。

体細胞WGS CNVの平滑化

セグメンテーションステージでは、同じコピー数が割り当てられ、同様の深度とBAFデータを持つ、隣接または近隣のセグメントが生成される場合があります。このセグメンテーションにより、一貫した真のコピー数をもつ領域がいくつかの断片に断片化される可能性があります。断片化は、コピー数推定の下流での使用に望ましくない場合があります。また、いくつかの用途では、真のコピー数変化またはアーティファクトに起因するかどうかにかかわらず、異なるコピー数が割り当てられる短いセグメントを平滑化することが好ましい場合があります。望ましくない断片化を低減するために、コール後のセグメント平滑化ステップ中に最初のセグメントをマージすることができます。

最初のコール後、`--cnv-filter-length`の指定値よりも短いセグメントは無視できると見なされます。残りの無視できないセグメントの間で、連続するペアがマージのために評価されます。試験的に、体細胞WGS CNV Callerは、互いの`--cnv-merge-distance`内にあり、同じCNおよびMCN割り当てを持つ連続する2つのセグメントを、間にある無視できるセグメントとともに、リコールされリスコアされる単一セグメントにします。マージされたセグメントが、その構成する、無視できない断片として同じCNおよびMCNを十分に高いクオリティスコアで受け取った場合、元のセグメントはマージされたセグメントで置き換えられます。マージされたセグメントは、さらに他の最初のセグメントまたはマージされたセグメントといずれかの側にマージされる場合があります。マージは、基準を満たすセグメントペアすべてがマージされるまで続行されます。

体細胞WGS CNV VCF出力

体細胞WGS CNV VCFファイルは、標準VCFフォーマットに従い、生殖細胞系列CNV VCF出力と以下の相違点があります。

ヘッダー

次のヘッダー行は、体細胞WGS CNVコールに固有です。

- **ModelSoure**：最終的な腫瘍モデルが選択された一次基準。次の値が有効です。
  - **DEPTH+BAF**：深度+BAFシグナルを用いて腫瘍モデルを決定します。

  - **DEPTH+BAF_DOUBLED**：初期深度+BAFモデルは、VAFシグナル、または予測深度変化の半分のところでの超過セグメントに基づいて複製されます。
  - **DEPTH+BAF_DEDUPLICATED**：深度+BAFモデルは、VAFシグナルまたは重複をサポートする不十分なセグメントに基づいて重複除去されます。
  - **DEPTH+BAF_WEAK**：深度+BAFシグナルを用いて、より低い信頼度の腫瘍モデルを決定します。
  - **VAF**：VAFシグナルは、不十分な深度+BAFシグナルによる腫瘍モデルを決定するために用いられます。
  - **DEGENERATE_DIPLOID**：サンプルは、深度+BAFおよびVAFからの適切なシグナルがない場合に、高純度二倍体として処理されます。二倍体カバレッジは、BAF = 50%のセグメント中のかなりの数の塩基で観察される最低値に設定されます。全VCFレコードで、`lowModelConfidence`が`FILTER`値に追加されます。
  - **SAMPLE_MEDIAN**：サンプルは、深度+BAFおよびVAFからの適切なシグナルがない場合に、高純度二倍体として処理されます。二倍体カバレッジはサンプル中央値に設定されます。全VCFレコードで、`lowModelConfidence`が`FILTER`値に追加されます。

- **EstimatedTumorPurity**：腫瘍によるサンプル中の推定細胞分画です。信頼できるモデルを決定できなかった場合、このフィールドの範囲は[0, 1]または`NA`です。
- **DiploidCoverage**：二倍体領域内のターゲットbinの予測リードカウントです。この数値に制限はありません。
- **OverallPloidy**：PASSイベントに関する腫瘍コピー数の長さ加重平均です。この数値に制限はありません。
- **AlternativeModelDedup**：全ゲノム重複を1つ減らすことに相当する最良のモデルに代わるものです。選択肢は一対の値（腫瘍の純度、二倍体カバレッジ）として与えられます。これは、最良のモデルが偽のゲノム重複を含む可能性がある場合の手作業による調査に有用です。
- **AlternativeModelDup**：1つ以上の全ゲノム重複に相当する最良のモデルに代わるものです。選択肢は一対の値（腫瘍の純度、二倍体カバレッジ）として与えられます。これは、最良のモデルが真のゲノム重複を見逃した可能性のある場合の手作業による調査に有用です。
- **OutlierBafFraction**：BAFが属するセグメントと適合しないb-アリル頻度の割合を測定するQCメトリクスです。高値の場合、ミスマッチした正常な、クロスサンプルコンタミネーション、または骨髄移植などの異なるモザイクゲノム源を示している可能性があります。このフィールドの範囲は[0,1]です。

ID

`ID`列はイベントのタイプを表します。GAIN、LOSS、およびREFイベントを表すことに加えて、コピーニュートラルなヘテロ接合性欠失（CNLOH）およびヘテロ接合性欠失を伴うコピー数増加（GAINLOH）エントリーは、ヘテロ接合性欠失（LOH）イベントを表します。

ALT

ALTフィールドは<DEL>、<DUP>のような2つのアリルを含むことができ、アリルのコピー数状態が異なる場合は、アリル特異的なコピー数を表現することができます。

FILTER

FILTERフィールドには、次の追加フィルターが適用されます。

- **binCount**：閾値よりも低いbinカウントでCNVイベントをフィルタリングします。
- **lowModelConfidence**：モデル推定値の信頼度を低くして、レコードをnon-PASSINGとしてマークします。

FORMAT

FORMATフィールドについては、ヘッダーセクションで説明します。以下のフィールドは体細胞WGS CNVに固有です。

ID	説明
AS	アリルリードカウント部位の数
BC	リードカウントbinの数
CN	サンプルの腫瘍画分中の総コピー数推定値
CNF	腫瘍コピー数の浮動小数点推定値
CNQ	正確な総コピー数Qスコア
MAF	マイナーアリル頻度の推定最大値
MCN	マイナーハプロタイプコピー数推定値
MCNF	腫瘍マイナーハプロタイプコピー数の浮動小数点推定値
MCNQ	マイナーコピー数Qスコア
NCN	正常サンプルコピー数。このフィールドは、Germline-awareモードでのみ表示されます。
SCND	CNとGCNの違い。このフィールドは、Germline-awareモードでのみ表示されます。
SD	セグメントに対するバイアス補正されたリードカウントの最良推定値

アリル特異的コピー数の例

Somatic WGS CNV Callerは、腫瘍の純度を推定することにより、腫瘍の総コピー数をレポートすることができます。マッチする正常なSNVまたは集団SNPからのBAF推定は、アリル特異的コピー数コールを可能にします。

下表に、リファレンス二倍体領域におけるDUPの例を示します。

合計コピー数	マイナーコピー数	ASCN シナリオ
4	2	2+2
4	1	3+1
[LOH] 4*	0	4+0

*このエントリーはヘテロ接合性欠失（LOH）のケースを表します。総コピー数は依然としてDUPと考えられているため、このエントリーには`GAINLOH`とアノテーションが付けられ、2+0とアノテーションが付けられるコピーニュートラルLOH`(CNLOH)`と区別されます。

ExpansionHunterを用いたリピート伸長の検出

Short tandem repeat（STR）は、リピート単位と呼ばれる短いDNAセグメントの反復からなるゲノム領域です。STRは正常な範囲を超えて伸長し、リピート伸長と呼ばれる変異を引き起こすことがあります。リピート伸長は、脆弱X症候群、筋萎縮性側索硬化症、ハンチントン病など多くの疾患の原因となっています。

DRAGENには、ExpansionHunterというリピート伸長検出メソッドがあります。ExpansionHunterは、各ターゲットリピートの内部および周辺に由来するリードのシーケンス-グラフベースの再アライメントを行います。次にExpansionHunterは、これらのグラフのアライメントに基づいて、各アリルのリピート長をジェノタイピングします。

ExpansionHunterは、PCRフリーの全ゲノムサンプル用に設計されています。座位のカバレッジが少なくとも10xである場合にのみ、リピートの遺伝型が決定されます。ExpansionHunterは、`fastq_list.csv`ファイル内の異なるライブラリーIDに割り当てられた複数のFASTQファイルでは実行することができません。

情報および解析について詳しくは、以下のExpansionHunterについての論文で入手可能です：

- Dolzhenko et al., *Detection of long repeat expansions from PCR-free whole-genome sequence data* 2017
- Dolzhenko et al., *ExpansionHunter:A sequence-graph based tool to analyze variation in short tandem repeat regions* 2019

リピート伸長検出オプション

DRAGENのリピート伸長検出を有効にするには、次のコマンドラインオプションが必要です。

- `--repeat-genotype-enable=true`
- `--repeat-genotype-specs=<path to specification file>`

`--sample-sex`オプションを使用して、サンプルの性別を指定できます。

次のオプションは任意です。

- `--repeat-genotype-region-extension-length=`<調査対象のリピート付近の領域の長さ>（初期設定値は1000 bp）

- `--repeat-genotype-min-baseq=<高信頼度の塩基の最低塩基クオリティ>` （初期設定値は20）

`---repeat-genotype-specs`オプションで指定される仕様ファイルについて詳しくは、177 ページの「リピート伸長仕様ファイル」 を参照してください。

リピート伸長検出の主な出力は、この解析によって検出されたバリアントを含むVCFファイルです。

リピート伸長仕様ファイル

リピート仕様（バリアントカタログとも呼ばれる）JSONファイルは、ExpansionHunterが解析するリピート領域を定義します。いくつかの病原性リピートのリピート仕様の初期設定は、DRAGENで使用されるリファレンスゲノムに基づいて、`/opt/edico/repeat-specs/`ディレクトリにあります。

提供される仕様ファイルの1つをテンプレートとして使用して、新たなリピート領域の仕様ファイルを作成できます。フォーマットの詳細については、ExpansionHunterの資料を参照してください。

`--repeat-genotype-specs`はExpansionHunterに必要です。このオプションを指定しない場合、DRAGENは、提供されたリファレンスに基づいて、`/opt/edico/repeat-specs/`から該当するカタログファイルを自動検出しようとします。

カバーされるリピート領域

最新のバリアントカタログには、AFF2、AR、ATN1、ATXN10、ATXN1、ATXN2、ATXN3、PHOX2B、ATXN7、ATXN8OS、C9ORF72、CACNA1A、CBL、CNBP、CSTB、DIP2B、DMPK、FMR1、FXN、HTT、JPH3、NOP56、PPP2R2B、TBP、TCF4、NIPA1、GLS、RFC1、およびPABPN1遺伝子に位置する疾患原因リピートに関する仕様が記載されています。また、GRCh38/hg38カタログは、NOTCH2NLリピートを定義します。このリピートは、この領域でのアライメントに問題があるため、GRCh37/hg19カタログには含まれていません。

ExpansionHunterは、FXN、ATXN3、ATN1、AR、DMPK、HTT、FMR1、ATXN1、C9ORF72リピートの病原性伸長を高い精度で検出することができます（上記のExpansionHunterの論文を参照）。いくつかのリピートの病原性の状態は、ExpansionHunterがコールしないシーケンスの中断やモチーフの変化の存在に依存する可能性があります。関連するリードアライメントを視覚的に確認する場合は、サードパーティー製のRepeat Expansion Viewerツールを使用できます。

リピート伸長検出出力ファイル

VCF出力ファイル

リピートジェノタイピングの結果は別個のVCFファイルとして出力され、これは、リピート仕様カタログファイルで定義された、コール可能なリピートそれぞれの各アリルの長さを提供します。名前は`<outputPrefix>.repeats.vcf (*.gz)`です。

VCF出力ファイルには、次のフィールドが最初に表示されます。

表 4　Core VCFフィールド

フィールド	説明
CHROM	染色体識別子
POS	リファレンスにおけるリピート領域前の最初の塩基位置
ID	Always .
REF	位置POSにおけるリファレンス塩基
ALT	<STRn>フォーマットのリピートアリルのリスト。Nはリピート単位の数です。
QUAL	Always .
FILTER	LowDepthフィルターは、座位全体の深度が10倍未満であるか、ブレークエンドの一方または両方にスパンするリード数が5未満である場合に適用されます。

表 5　追加のINFOフィールド

フィールド	説明
END	リファレンス内のリピート領域の最後の塩基位置
REF	リファレンス内のリピートがスパンするリピートユニット数
RL	リファレンスの長さ（単位：bp）
VARID	バリアントカタログからのバリアントID
RU	リファレンス方向のリピートユニット
REPID	バリアントカタログからのバリアントID

表 6　GENOTYPE（サンプルあたり）フィールド

フィールド	説明
GT	遺伝型
SO	アリルを支持するリードのタイプ。値は、SPANNING、FLANKING、またはINREPEATの可能性があります。これらの値は、リードがスパン、フランク、または完全にリピート内に含まれるかどうかを示します。
REPCN	アリルがスパンするリピートユニット数
REPCI	REPCNの信頼区間
ADSP	アリルと一致するスパンニングリード数
ADFL	アリルと一致するフランキングリード数
ADIR	アリルと一致するリピート内リード数
LC	座位カバレッジ

例えば、以下のVCFエントリーは、サンプルNA13537におけるATXN1リピートの記述です。

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA13537
chr6 16327864 . G <STR33>,<STR58> . PASS
END=16327954;REF=30;RL=90;RU=TGC;VARID=ATXN1;REPID=ATXN1
GT:SO:REPCN:REPCI:ADSP:ADFL:ADIR:LC 1/2:SPANNING/INREPEAT:33/58:33-33/52-
71:4/0:69/83:0/4:37.459459
```

この例では、第1のアリルは33個のリピートユニットをスパンし、第2のアリルは58個のリピートユニットをスパンしています。リピート単位はTGC（RU INFOフィールド）のため、最初のアリルのシーケンスはTGC x 33で、2番目のアリルのシーケンスはTGC x 58です。このリピートは、リファレンス内の30リピート単位にスパンします（REF INFOフィールド）。

短いアリルの長さはスパンニングリード（SPANNING）から推定し、アリルの長さはリピート内リード（INREPEAT）から推定しました。伸長したアリルの大きさの信頼区間は（52,71）です。サイズ33のリピートアリルと一致する4つのスパンニングリードおよび69のフランキングリードがあります。4つのリードはサイズ33のリピートを完全に含み、69のフランキングリードは最大33のリピートユニットで重複します。サイズ58のリピートアリルと一致する83のフランキングリードと4つのインリピートリードが存在します。この座位の平均カバレッジは37.46倍です。

その他の出力ファイル

ターゲットリピート領域におけるリードのシーケンスグラフアライメントは、BAMファイルに出力されます。GitHubにある専用のGraphAlignmentViewerツールを使用して、アライメントを視覚化することができます。Integrative Genomics Viewer（IGV）のようなプログラムは、グラフでアライメントしたリード値を表示するように設計されておらず、これらのBAMSを視覚化することはできません。

BAMでは、<LocusName>、<StartPosition>、<GraphCIGAR>のフォーマットを使用して、グラフアライメントをカスタムXGタグに保存します。

- **LocusName**：リピート伸長仕様ファイル内の対応するエントリーと一致する座位識別子です。
- **StartPosition**：最初のグラフノードでのリードの開始アライメント位置です。
- **GraphCIGAR**：その位置から始まるグラフに対するリードのアライメントです。GraphCIGARは、一連のグラフノード識別子と、各ノードへのリードのアライメントを記述する線形CIGARSから構成されます。

BAMファイル内のクオリティスコアはバイナリーです。高スコアの塩基には40点、低スコアの塩基には0点が割り当てられます。

脊髄性筋萎縮症コール

個体におけるSMN1遺伝子の全コピーの破壊は、脊髄性筋萎縮症（SMA）を引き起こします。SMN1は高い同一性のパラログであるSMN2を有し、約10個のSNVと小さなindelのみが異なります。これらのうちの1つ（hg19 chr5:70247773 C->T）はスプライシングに影響し、SMN2からの機能的SMNタンパク質の産生を大きく阻害します。標準的なWGS解析では、SMNの完全なバリアントコール結果は生成されません。これは、この高い類似性の重複と共通のコピー数バリエーションが組み合わさっているためです。しかし、SMAケースの約95%は、SMNの任意のコピーにおける機能性C（SMN1）アリルの欠如を決定することによって検出することができます。

DRAGEN SMAコールは、SMN1およびSMN2を表す単一のリファレンスへのリードをアライメントさせるために、シーケンス-グラフ再アライメントを使用します。標準的な二倍体ジェノタイプコールに加えて、DRAGENはCアリルの存在を調べるために直接統計検定を用います。Cアリルが検出されない場合、そのサンプルは影響を受けたサンプルと呼ばれ、そうでなければ影響を受けていないサンプルと呼ばれます。

SMAコールは、PCRフリーライブラリー中のヒト全ゲノムシーケンスサンプルに対してのみサポートされます。

個体におけるSMN1遺伝子の全コピーの破壊は、脊髄性筋萎縮症（SMA）を引き起こします。SMN1は高い同一性のパラログSMN2をもちます。SMN2は、約10個のSNVと小さなindelだけが異なります。例えば、hg19 chr5:70247773 C->Tはスプライシングに影響し、SMN2からの機能的SMNタンパク質の産生を大きく阻害します。共通コピー数バリエーションと組み合わされた高い類似性の重複のために、標準的な全ゲノムシーケンス（WGS）解析では、SMNについて完全なバリアントコールの結果は得られません。SMAケースの95%はSMN1のいずれかのコピーに機能性C（SMN1）アリルが存在しないことに起因するため、ターゲットコールソリューションはSMAの検出に有効な可能性があります。

DRAGENは、WGSデータからSMAのステータスを検出するために、以下の2つの独立したコンポーネントを提供します。

- ExpansionHunter
- SMN Caller

SMAコーラーは、SMAステータスに加えて、SMN1/SMN2コピー数をレポートし、SMAキャリアーステータスを識別します。

[1]Wirth B. An update of the mutation spectrum of the survival motor neuron gene (SMN1) in autosomal recessive spinal muscular atrophy (SMA). *Human Mutation*. 2000;15(3):228-237. doi:10.1002/(sici)1098-1004(200003)15:3<228::aid-humu3>3.0.co;2-9

ExpansionHunterによるSMAコール

SMAコールは、シーケンス-グラフ再アライメントを使用してリピート伸長検出とともに実行され、SMN1およびSMN2を表す単一のリファレンスにリードをアライメントします。

標準的な二倍体ジェノタイプコールに加えて、ExpansionHunterによるSMAコールでは、Cアリルの存在を調べるために直接統計検定を用います。Cアリルが検出されない場合、そのサンプルは影響を受けたサンプルと呼ばれ、そうでなければ影響を受けていないサンプルと呼ばれます。

SMAコールは、PCRフリーライブラリーを用いたヒト全ゲノムシーケンスに対してのみサポートされています。

リピート伸長検出とともにSMAコールを有効にするには、`--repeat-genotype-enable`オプションを`true`に設定します。グラフアライメントオプションについては、176 ページの 「ExpansionHunterを用いたリピート伸長の検出」 を参照してください。

SMAコールを有効にするには、バリアント仕様カタログファイルに、ターゲットのSMN1/SMN2バリアントを記述する必要があります。`/opt/edico/repeat-specs/experimental`フォルダーには、ファイルの例が含まれています。

`<outputPrefix>.repeat.vcf`ファイルには、SMN出力とターゲットリピートが含まれています。SMN出力は、SMN1のスプライスに影響する位置での単一のSNVコールとして表され、次のカスタムフィールドでSMAステータスが使用されます。

表 7　repeat.vcf出力ファイル内のSMA結果

フィールド	説明
VARID	SMNはSMNコールをマークします。
GT	正常な(二倍体)遺伝型モデルを用いた、この位置でのジェノタイプコールです。
DST	SMAステータスコール: +は検出されたことを示します -は検出されていないことを示します ?は未確定を示します
AD	CおよびTアリルをサポートする総リード数。
RPL	影響を受けていないモデルと影響を受けたモデルの間のlog10尤度比です。正のスコアは、影響を受けていないモデルの可能性が高いことを示します。

SMN Caller

SMN Callerは、SMN1およびSMN2のコピー数をコールすることによって、SMAステータスおよびSMAキャリアーステータスを検出します。このコーラーは、*脊髄性筋萎縮症診断およびゲノムシーケンスデータからのキャリアースクリーニングで実施したメソッドから得られます*[1]。

SMN Callerを有効にするには、`--enable-smn=true`を、生殖細胞系列のみのWGS解析ワークフローの一部として使用します。SMN Callerは初期設定で無効になっています。

SMN Callerは次のステップを実行します：

1. SMNの総コピー数とインタクトなコピー数の特定
2. 8つのdifferentiating siteでSMN1のコピー数をコール
3. differentiating siteコールからSMAとSMAのキャリアーステータスを判別

SMN Callerは、少なくとも30xカバレッジでヒトリファレンスゲノムにアライメントされたWGSデータを必要とします。

総SMNコピー数およびインタクトSMNコピー数

SMN1とSMN2における2つの共通コピー数バリアント（CNV）は全遺伝子CNVとエクソン7と8の部分的な遺伝子欠失を含みます。

SMN1またはSMN2のいずれかにアライメントしたリードがカウントされます。エクソン1からエクソン6までのリードカウントを用いて、総SMNコピー数を決定します。エクソン7および8におけるリードカウントを用いて、エクソン7および8の欠失を有さないSMNコピー（インタクトSMNコピー数）を決定します。これら2つの領域についてのSMNコピー数を推定するために、リードカウントを、ゲノム全体にわたる3,000の予め選択された2 kb領域から得られる二倍体ベースラインに対してノーマライズします。3,000のノーマライズ領域は、集団サンプル全体にわたって安定したカバレッジを有するリファレンスゲノムの部分からランダムに選択されます。次いで、SMN Callerは、総SMNコピー数からインタクトSMNコピー数を減算することによって、エクソン7および8の欠失を有するSMNコピーの数を計算します。

differentiating siteでのSMN1コピー数

SMN1コピー数を計算するために、コーラーはSMN1とSMN2のエクソン7と8にある既定の8つのdifferentiating siteを用います。これらの部位の1つは、ExpansionHunterとのSMAコールに使用されるスプライス部位バリアントです（180 ページの 「ExpansionHunterによるSMAコール」 を参照）。コーラーは、SMN1とSMN2との間にシーケンスの違いがある位置でdifferentiating siteを選択します。ここでSMN1コピー数をコールすることが、1,000ゲノムプロジェクトからのシーケンスデータに基づくと、最も正しいと思われます。

各differentiating siteについて、SMN1特異的アリルおよびSMN2特異的アリルを、SMN1またはSMN2中の相同領域のいずれかにマッピングするリードでカウントします。コーラーは前ステップで計算されたインタクトSMNコピー数を前提として、二項モデルを用いて、2つの遺伝子特異的なカウントから、それぞれの可能性のあるSMN1コピー数の尤度を計算します。

SMAおよびSMAのキャリアーステータス

DRAGENは、単一のSMN1コピー数を同定するために8つのdifferentiating siteを考慮した尤度モデルを使用します。モデルが単一のSMN1コピー数に対する有意な裏付けを示す場合、DRAGENはそのSMN1コピー数を用いてSMAステータスおよびキャリアーステータスを以下のように決定します。

SMN1 CN	SMAステータス	SMAキャリアーステータス
0	True	False
1	False	True
>1	False	False

尤度モデルが単一のSMN1コピー数に対する有意な裏付けを示さない場合、differentiating siteは、SMN1コピー数が2未満であるという仮説に反しているとみなされます。コピー数が2以上であれば、SMAステータスおよびSMAキャリアーステータスは共に`False`です。

単一の信頼できるSMN1コピー数が存在せず、かつSMN1コピー数が確信を持って2以上でない場合、スプライス部位を用いてSMAステータスをコールします。このコーラーはスプライス部位のアリルカウントを用いてSMAステータスを以下のように決定します。

スプライス部位機能性C（SMN1）アリルが欠如している証明	SMA ステータス	SMAキャリアース テータス
強い	True	False
弱い	None	None
強く否定	False	None

SMNコマンドラインの例

SMN Callerを有効にするには、`--enable-smn=true`を使用します。SMN Callerは初期設定で無効になっています。SMN Callerは、マッパーを有効にしたFASTQインプットから、または事前アライメント済みのBAM/CRAMインプットから実行できます。SMN Callerは、WGS生殖細胞系列解析ワークフローの他の生殖細胞系列バリアントコーラーと並行して有効にすることもできます。その他のバリアントコーラーについては、65 ページの 「DRAGEN DNA Pipeline」 を参照してください。

FASTQインプット

次のコマンドラインの例では、FASTQインプットを使用しています。

```
dragen \
-r /staging/human/reference/hg38_alt_aware/DRAGEN/8 \
--fastq-file1 /staging/test/data/NA12878_R1.fastq \
--fastq-file2 /staging/test/data/NA12878_R2.fastq \
--output-directory /staging/test/output \
--output-file-prefix NA12878_dragen \
--RGID DRAGEN_RGID \
--RGSM NA12878 \
--enable-map-align=true \
--enable-smn=true
```

事前アライメント済みのBAMインプット

次のコマンドラインの例では、アライメント済みのBAMインプットを使用しています。

```
dragen \
-r /staging/human/reference/hg38_alt_aware/DRAGEN/8 \
--bam-input /staging/test/data/NA12878.bam \
--output-directory /staging/test/output \
--output-file-prefix NA12878_dragen \
--enable-map-align=false \
--enable-smn=true
```

SMN出力ファイル

SMN Callerは、出力ディレクトリに`<output-file-prefix>.smn.tsv`ファイルを生成します。出力ファイルには、ヘッダー行と、次のタブ区切りフィールドを含むサンプルコール行が含まれます。

フィールドヘッダー	説明	値
サンプル	サンプル名	文字列

フィールドヘッダー	説明	値
isSMA	SMAの影響ステータス	• True • False • None*
isCarrier	SMAキャリアーステータス	• True • False • None*
SMN1_CN	SMN1のコピー数	• 負でない整数 • None*
SMN2_CN	SMN2のコピー数	• 負でない整数 • None*
SMN2delta7-8_CN	SMN2Δ7-8のコピー数（エクソン7と8の欠失）	負でない整数
Total_CN_raw	総SMNの処理前ノーマライズ深度	浮動小数点
Full_length_CN_raw	インタクトSMNの処理前ノーマライズ深度	浮動小数点
SMN1_CN_raw	differentiating siteの処理前のSMN1 CN値	カンマ区切りの8つの浮動小数点値

*値`None`は、確信のあるコールを行えなかったことを示します。

`*.tsv`出力ファイルの例を次に示します。

```
#Sample isSMA isCarrier SMN1_CN SMN2_CN SMN2delta7-8_CN Total_CN_raw Full_
length_CN_raw SMN1_CN_raw
HG00111 False False 2 3 2 4.88 5.04 1.00,2.74,2.10,1.89,2.21,1.63,2.47,2.00
```

[1]Chen X, Sanchis-Juan A, French CE, et al. Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. *Genetics in Medicine.* 2020;22(5):945-953. doi:10.1038/s41436-020-0754-0

CYP2D6 Caller

CYP2D6 Callerは、全ゲノムシーケンス（WGS）データからCYP2D6遺伝子のジェノタイピングを行うことができるもので、Cyrius[1]で実施したメソッドに基づきます。偽遺伝子パラログCYP2D7との高いシーケンス類似性と広範囲の共通な構造多型（SV）のため、バリアントを解明し、適当なスターアリルハプロタイプを同定するために特殊なコーラーが必要です。

CYP2D6 Callerは次のステップを実施します：

1. リード深度からCYP2D6およびCYP2D7の総コピー数を決定します。

2. CYP2D6/CYP2D7のdifferentiating siteにおけるCYP2D6由来のコピー数を決定します。
3. CYP2D6遺伝子に沿ったCYP2D6由来コピー数の変化を計算することにより、SVブレークポイントを検出します。
4. CYP2D6コピーのスモールバリアントコールを行います。
5. 検出されたSVブレークポイントとスモールバリアントからスターアリルを同定します。
6. スターアリルと呼ばれるものの中で最も可能性の高い遺伝型を同定します。

CYP2D6 Callerは、少なくとも30xカバレッジでヒトリファレンスゲノムにアライメントされたWGSデータを必要とします。

CYP2D6およびCYP2D7の総コピー数

CYP2D6コールの第1ステップは、CYP2D6とCYP2D7の組み合わせコピー数を決定することです。CYP2D6またはCYP2D7のいずれかの領域にアライメントしたリードをカウントします。各領域のカウントをGCバイアスで補正し、次に二倍体ベースラインに対してノーマライズします。GCバイアス補正およびノーマライゼーション因子は、ゲノム全体にわたる3,000個の予め選択された2 kb領域におけるリードカウントから決定されます。次いで、CYP2D6およびCYP2D7領域にわたる平均シーケンス深度から、CYP2D6およびCYP2D7の組み合わせコピー数を計算します。

differentiating site

CYP2D6由来のコピー数は、CYP2D6遺伝子全体の117の既定のdifferentiating siteで計算されます。differentiating siteは、CYP2D6およびCYP2D7のシーケンスが異なる位置で選択され、1,000ゲノムプロジェクトからのシーケンスデータに基づいて、CYP2D6由来のコピー数をコールし、98%を上回る精度を示します。

各differentiating siteについて、CYP2D6特異的アリルおよびCYP2D7特異的アリルを、CYP2D6またはCYP2D7の相同領域のいずれかにマッピングするリードでカウントします。次いで、CYP2D6由来のコピー数を、前ステップから計算したCYP2D6とCYP2D7の総コピー数を用いて、2つの遺伝子特異的アリルカウントから計算します。

構造多型コール

CYP2D6遺伝子に沿ったCYP2D6由来のコピー数は、遺伝子全体の欠失および重複、ならびに特定の遺伝子変換および遺伝子融合など、既知の集団構造多型（SV）を同定するために用いられます。以下の融合バリアントが検出されます：

融合ブレークポイント	ハイブリッド遺伝子の構造	スターアリルの指定
エクソン9	2D6-2D7	*4.013、 *36、 *57、 *83

融合ブレークポイント	ハイブリッド遺伝子の構造	スターアリルの指定
エクソン9	2D7-2D6	*13
イントロン4	2D7-2D6	*13
イントロン1	2D7-2D6	*13
イントロン1	2D6-2D7	*68

エクソン9融合ブレークポイントに加えて、エクソン9はCYP2D7遺伝子変換に関与することができ、その結果、真のハイブリッドの代わりに埋め込みCYP2D7シーケンスが生じます。構造多型コーラーはエクソン9の遺伝子変換も検出します。CYP2D6由来のコピー数の変化のみが構造多型を生じさせるため、2つのハイブリッドコピーによって構造多型コールを生じさせないまれなケースになる可能性があります。例えば、エクソン9に融合ブレークポイントをもつ*36と*13の両方が存在する場合です。しかし、構造多型コーラーは、同じ融合型の複数のコピー（例：*36x2）を検出することができ、また、エクソン9の遺伝子変換コピーとエクソン9 2D6-2D7のハイブリッドの両方が存在する場合も検出できます。

スモールバリアントコール

さまざまなスターアリルを定義する118のスモールバリアントをリードアライメントから検出します。これらのバリアントのうち96は、CYP2D6の固有の（非相同）領域に存在し、高いマッピングクオリティを有します。CYP2D6にマッピングするリードのみが、非相同領域におけるバリアントのコールに用いられます。他の22のバリアントはCYP2D6の相同領域に存在し、CYP2D6またはCYP2D7のいずれかにマッピングするリードがバリアントコールに用いられます。

各バリアントについて、バリアントアリルまたは非バリアントアリルのいずれかを含むリードをカウントします。次いで、シーケンス誤差を組み込んだ二項モデルを用いて、最も可能性の高いバリアントコピー数（非バリアントは0）を決定します。ストランドバイアスフィルターは、偽陽性のコールを持つ傾向がある特定のノイズの多いバリアントに適用されます。

シーケンスのクオリティが低いサンプル、またはCYP2D6のコピー数が5を超えるサンプルでは、アリルカウントのばらつきが大きくなります。このばらつきが大きいと、最も可能性の高いバリアントのコピー数が間違っている可能性が高まります。こうした状況に対処するために、スモールバリアントコーラーは、可能性の低い別のバリアントのコピー数も示します。

スターアリルの同定

SVと呼ばれる遺伝型とスモールバリアントジェノタイプは、128種類の異なるスターアリルの定義とマッチします。この結果、バリアントジェノタイプと呼ばれる遺伝型と一致するスターアリルの異なるセットが生じる可能性があります。例えば、*1、*46、*43、*45などでは、両方のスターアリルのセットが同じ4つのスモールバリアントを含みます。スモールバリアントコーラーが、最も可能性の高いバリアントのコピー数に加えて、可能性の低いバリアントのコピー数を代替として放出する場合、これらのバリアントのコピー数の代替セットもスターアリルの定義とマッチするため、結果として、スターアリルの異なるセットが同定される可能性があります。マッチするスターアリルの数は、前ステップで決定されたCYP2D6由来の遺伝子コピーの数と一致する必要があります。CYP2D6由来の遺伝子コピーが2つ未満の場合、1つ以上の*5欠失ハプロタイプがスターアリルの出力セットに含まれます。すべてのバリアントジェノタイプがスターアリルのセットにマッチできない場合、CYP2D6 Callerはジェノタイピングステップ中にno callをフィルター値`No_call`で返します。

ジェノタイピング

可能性のあるスターアリルのセットを前提として、ジェノタイピングステップは、セット中のスターアリルすべてを含む2つの可能性のあるハプロタイプを同定しようとします。欠失ハプロタイプ（*5）はこのプロセスでハプロタイプの可能性があると考えられます。与えられた遺伝型の尤度は、1,000ゲノムプロジェクトで決定された集団内頻度の表から決定され、最も高い集団内頻度をもつ遺伝型が選択されます。2つ以上の可能性のある遺伝型が同じような集団内頻度で同定された場合、遺伝型すべてが放出されます。この結果、フィルター値`More_than_one_possible_genotype`のコールが行われます。

CYP2D6出力ファイル

CYP2D6 Callerは、出力ディレクトリに`<output-file-prefix>.cyp2d6.tsv`ファイルを生成します。この出力ファイルには、ヘッダー行のない次のタブ区切りフィールドが含まれます：

- サンプル名。
- セミコロンで区切られた1つ以上のCYP2D6遺伝型、またはno callが`None`。
- フィルターのステータス。値には次のものが含まれます：`PASS`、`No_call`、または`More_than_one_possible_genotype`。

それぞれのCYP2D6遺伝型は、スラッシュで区切られた2つのハプロタイプ（例：*1/*2）を含みます。各ハプロタイプは、正符号（例：*10+*36）で分離された1つ以上のスターアリルからなります。ハプロタイプに同じスターアリルが1コピー以上ある場合、そのスターアリルは1回だけ出現し、その後に、乗法記号、さらにコピーの数が続きます（例：*1の2コピーは*1x2）。

出力ファイルの例

出力ファイルの例を次に示します：

```
NA18632 *10+*36x2/*52 PASS
HG01190 *4+*68/*5 PASS
NA17244 *2/*4x2+*13+*83;*2/*4+*4.013+*13+*39 More_than_one_possible_
genotype
NA19908 *1/*46;*43/*45 More_than_one_possible_genotype
NA18611 None No_call
```

コマンドラインの例

CYP2D6 Callerを有効にするには、`--enable-CYP2D6=true`を使用します。CYP2D6 Callerは初期設定で無効になっています。CYP2D6 Callerは、マッパーを使用してFASTQインプットから直接実行することも、事前にアライメントされたBAM/CRAMインプットから実行することもできます。WGS生殖細胞系列解析ワークフローの一部として、CYP2D6 Callerを他の生殖細胞系列バリアントコーラーと並行して有効にすることもできます。バリアントコーラーの詳細については、65 ページの 「DRAGEN DNA Pipeline」 を参照してください。

FASTQインプット

次のコマンドラインの例では、FASTQインプットを使用しています。

```
dragen \
-r /staging/human/reference/hg38_alt_aware/DRAGEN/8 \
--fastq-file1 /staging/test/data/NA12878_R1.fastq \
--fastq-file2 /staging/test/data/NA12878_R2.fastq \
--output-directory /staging/test/output \
--output-file-prefix NA12878_dragen \
--RGID DRAGEN_RGID \
--RGSM NA12878 \
--enable-map-align=true \
--enable-cyp2d6=true
```

事前アライメント済みのBAMインプット

次のコマンドラインの例では、アライメント済みのBAMインプットを使用しています。

```
dragen \
-r /staging/human/reference/hg38_alt_aware/DRAGEN/8 \
--bam-input /staging/test/data/NA12878.bam \
--output-directory /staging/test/output \
--output-file-prefix NA12878_dragen \
--enable-map-align=false \
--enable-cyp2d6=true
```

[1]Chen X, Shen F, Gonzaludo N, et al. Cyrius: accurate CYP2D6 genotyping using whole-genome sequencing data. *The Pharmacogenomics Journal.* 2021;21(2):251-261. doi:10.1038/s41397-020-00205-5

構造多型コール

DRAGEN構造多型（SV）コーラーは、SVを提供するためにManta構造多型コールを統合し、拡張し、50塩基以上のSVおよびindelのコールを提供します。SVとIndelはマップされたペアエンドシーケンスリードからコールされます。SVコーラーは、少数の個体セットにおける二倍体生殖細胞系列バリエーションの解析に最適化されています。

SVコーラーは、次のアクションを実行します：

- 1つの効率的なワークフロー内で、大規模なSV、中型のindel、大型の挿入を発見、アセンブリ、スコアリングします。

- 精度を向上させるために、SV発見時およびスコアリング時にペアリードとスプリットリードの証拠を組み合わせます。ただし、強力な証拠がある場合にバリアントをレポートするために、スプリットリードや成功したブレークポイントアセンブリを必要としません。
- 1つまたは複数のインプットサンプルに対して、インプットVCFファイルからの既知のSV欠失および挿入を、スタンドアロンの手順として、または標準的なSV発見とともにスコアリングします。
- 二倍体サンプルの小さなセットにおける生殖細胞系列バリアントのスコアリングモデルを提供します。

SVとindelの推定はすべて、VCF 4.1フォーマットで出力されます。

DRAGEN SV Callerの概要

DRAGEN SV Callerは、SVとindelの発見プロセスを以下のステップに分けます。

1. 入力ファイルを読み込み、断片サイズ分布や染色体レベル深度などのアライメント統計を推定します。SV Callerのインプットオプションについて詳しくは、198 ページの 「コマンドラインオプション」 を参照してください。
2. ゲノムをスキャンして、すべてのSV関連領域のブレークエンド関連グラフなど、ゲノム全体にわたるさまざまなデータ構造を構築します。このグラフには、ブレークエンド関連の可能性があるゲノムの領域すべてを連結するエッジが含まれます。エッジはゲノムの2つの異なる領域を連結して長距離結合の証拠を示すことができ、あるいはエッジは局所的なindel/スモールSV関連を捕捉する領域に連結することができます。これらの関連は、特定のSV仮説よりも一般的であり、複数のブレークエンド候補が1つのエッジ上で見つかる可能性があります。通常、エッジごとに1つまたは2つの候補のみが検出されます。
3. ブレークエンド関連グラフを解析して候補SVを発見し、発見された候補SVおよびインプットからの既知のSVをスコアリングします。解析とスコアリングは次のように行われます。
   a. 指定されたグラフエッジに関連付けられているSV候補を推定します。
   b. SVブレークエンドをアセンブルします。
   c. 発見されたSV候補を、インプットデータに含まれる既知のSV候補とマージします。
   d. さまざまな生物学的モデル（現在は生殖細胞系列および体細胞系）の下で、各SV候補をスコアリング/ジェノタイピングおよびフィルタリングします。
   e. スコアリングされたSVをVCFに出力します。

DRAGEN SV Callerの機能

DRAGEN SV Callerは、コピー数解析や大規模な*de novo*アセンブルを行わなくても同定可能な構造多型のタイプすべてを発見することができます。検出可能なタイプの詳細については、191 ページの 「検出されたバリアントクラス」 を参照してください。

各構造多型およびindelについて、SV Callerは、ブレークエンドを塩基対解像度にアセンブルし、左シフトブレークエンド座標を（VCF 4.1 SVレポート作成ガイドラインに従って）、ブレークエンド間のブレークエンド相同性シーケンスおよび/または挿入シーケンスとともにレポートしようとします。アセンブリ結果について信頼性できるデータ説明を提供できないことがよくあります。この場合、バリアントはIMPRECISEとしてレポートされ、ペアエンドリードの証拠のみに従ってスコアリングされます。

既知のSVを強制ジェノタイピングのインプットとして提供することができます。この既知のSVインプットは、スタンドアロンで、または標準のSV発見ワークフローと一緒にスコアリングすることができます。この場合、既知のSVと発見されたSVはマージされます。

SV Callerへのインプットとして提供されるシーケンスリードは、各シーケンス断片の2つのリードの間に「innie」配向を生じ、それぞれが断片挿入物の外端から内側へのリードを提示する、ペアエンドシーケンスアッセイからのものであることが期待されます。

SV Callerは全ゲノムおよびDNA上の全エクソーム（またはその他のターゲット濃縮）シーケンスアッセイ用に主にテストされます。これらのアッセイでは、以下のアプリケーションがサポートされます：

- 5以下の二倍体個体のジョイント解析
- マッチするTumor-Normaサンプルペアの減法解析
- 個々の腫瘍サンプルの解析

ジョイント解析では、より大きなコホートに対する特定の制約はないものの、安定性またはコールクオリティの問題がある可能性があります。

血液腫瘍サンプルに対して体細胞コールを行う場合、マッチする正常サンプルは腫瘍細胞でコンタミネーションされている可能性があります。コンタミネーションは、体細胞バリアントのコール率を大幅に低下させることができます。Tumor-in-Normal（TiN）コンタミネーションを説明するには、血液腫瘍モードを有効にします。詳細については、198 ページの 「血液腫瘍コール」 を参照してください。

腫瘍サンプルは、マッチする正常サンプルなしで解析することができます。この場合、スコアリング機能は利用できませんが、裏付けとなる証拠の数は利用可能であり、多くのフィルターを有効に適用することができます。

操作モード

構造多型コールは次のモードで実行できます：

- スタンドアロン：マップされたBAM/CRAM入力ファイルを使用します。データのマッピングとアライメントをまだ行っていない場合は、194 ページの 「インプット要件」 を参照してください。このモードには、次のオプションが必要です：
  - *--enable-map-align false*
  - *--enable-sv true*
- 統合：DRAGENマッパー/アライナーの出力で自動的に実行されます。このモードには、次のオプションが必要です：
  - *--enable-map-align true*
  - *--enable-sv true*
  - *--enable-map-align-output true*
  - *--output-format bam*

また、他の任意のコーラーとの構造多型コールを有効にすることもできます。

統合モードのコマンドラインの例を次に示します：

```
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>
```

スタンドアロンモードでのジョイント二倍体コールのコマンドラインの例を次に示します：

```
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>
```

検出されたバリアントクラス

SV Callerは、ゲノム中の新規DNA隣接関係として説明できるバリエーションクラスをすべて発見することができます。新規DNA隣接関係はブレークエンドパターンに基づいて以下のカテゴリーに分類されます：

- 欠失
- 挿入。SV挿入は、挿入されたシーケンスが完全にアセンブルされるかどうかによって、以下の2つのサブクラスに分けることができます。
  - 完全にアセンブルされた挿入
  - 部分的にアセンブルされた（すなわち、推定された）挿入
- タンデム重複
- 染色体内および染色体間の転座、または複雑な構造多型に対応する未分類のブレークエンドペア。

既知の制限事項

SV Callerは、次のバリアントタイプを直接検出できません：

- 分散重複
  分散重複は、挿入あるいは未分類ブレークエンドと間接的に呼ぶことができます。

- リファレンス縦列リピートの大半の伸長/収縮バリアント。
- 小さな逆位に相当するブレークエンド。
  - 制限されるサイズは調べられていませんが、理論的には約200塩基未満で検出が低減します。マイクロインバージョンは、挿入/欠失を組み合わせたバリアントとして間接的に検出される可能性があります。
- 完全にアセンブルされた大きな挿入。
  - 完全にアセンブルされた挿入断片の最大サイズは、ペアリードの断片サイズの約2倍に相当しますが、挿入断片を完全にアセンブルする能力は、このサイズ以前では実用的でないレベルまで低下します。
  - SV Callerは、そのようなイベントのブレークエンドサインが見つかると、挿入されたシーケンスが完全にアセンブルできなくても、極めて大きな挿入を検出し、レポートします。

より一般的なリピートベースの制限は、全バリアントタイプにあります：

- バリアントをブレークエンド解像度にアセンブルする力は、ブレークエンドのリピート長がリードサイズに近づくにつれて0に低下します。
- ブレークエンドを検出する力は、ブレークエンドリピート長が断片サイズに近づくにつれて、（ほぼ）0に低下します。

SV Callerはある種の新しいDNA隣接関係をバリアントクラスに分類しますが、複雑な再編成から生じる高レベルのイベントを推定する能力には限界があります。そのため、欠失、重複、挿入として要約されるある種のコールは、ブレークエンドの完全なシステムと特定のイベントに関連するコピー数の変化を見ることによって、より適切に記述できる可能性があります。

強制ジェノタイピング機能

DRAGEN SV Callerは、VCFファイルからインプットされたSVのセットを強制ジェノタイピングすることができます。強制ジェノタイピングとは、サンプルデータでバリアントがサポートされていない場合でも、インプットされたSVがスコアリングされ、SV Callerの出力に放出されることを意味します。例えば、生殖細胞系列解析の場合、バリアントのクオリティはSVがコールされるために通常必要な閾値を下回ったとしても、インプットバリアントは処理され、出力VCFに書き込まれます。

強制ジェノタイピングは通常、既知のSVを標準的なSV発見よりも高いコール率で検出することを可能にします（特に低深度サンプルでのSV発見の場合）。強制ジェノタイピングは、SVアリルの存在を否定するためにも有用です。例えば、強制ジェノタイピングを用いて、確信のあるホモ接合性リファレンス遺伝型と、SV座位のシーケンスカバレッジの欠如とを区別することができます。

強制ジェノタイピングSVは、現在実行されているSV解析に従って処理されます。例えば、1つ以上の正常サンプルをインプットとして提供することによって生殖細胞系列解析が構成される場合、インプットSVは生殖細胞系列モデルの下でスコアリングされます。

強制ジェノタイピングアリルは常に出力に放出され、サンプルデータからのみ発見されたSVと比較して、スコアリングルールが修正され、フィルタリングルールが適用されている可能性があります。

強制ジェノタイピングモード

強制ジェノタイピングは、2つのモードで実行することができます。

- スタンドアロン：インプットVCFに記述されているSVだけがスコアリングされ、放出されます。
- 統合：標準のSV解析が実行され、結果は強制ジェノタイピングのインプットからスコアリングされたSVと統合されます。このワークフローでは、サンプルデータから検出されたSVと強制ジェノタイピングアリルの組み合わせを出力します。このワークフローは、`--sv-discovery`オプションがtrueの場合は常に実行されます。

強制ジェノタイピングのインプット

*--sv-forcegt-vcf*オプションを使用して、強制ジェノタイピングのインプットを指定できます。インプットはSVアリルのVCFである必要があります。SVアリルタイプは挿入、欠失、タンデム重複、ブレークエンドに限定されており、`INFO/UNINTENSIVE`フラグで標識されません。以下は、VCFレコードをインプットSVアリルとして処理するために必要なフィルタリング基準です。この基準のいずれかが満たされない場合、VCFレコードは、強制ジェノタイピングのためのインプットSVのセットから除去されます。強制ジェノタイピングVCFがコマンドラインで指定された場合、SV Callerは、インプットSVとして使用されたSVレコードの総数と、以下の基準によりフィルタリングされたレコード（ある場合）の総数をレポートします。

- 挿入、削除、タンデム重複、またはブレークエンドレコードを記述します。
- `INFO/INCORRECT`フラグを含めることはできません。
- 複数のALTアリルを含むことはできません。
- `PASS`または不明（.）の`FILTER`値を持ちます。
- indelはすべて、最小スコア付きバリアントサイズ以上です（初期設定値は50）。
- 同じファイルにすでに記述されているSVアリルをリピートすることはできません。
- `REF`フィールドを空または不明（.）にすることはできません。

完全な挿入シーケンスを記述する`ALT`エントリーを含め、VCFの小さなindelフォーマットを使用して挿入を記述する必要があります。ALTアリルの記号として`<INS>`を用いることは認められません。欠失は、VCFの小さなindelフォーマットまたは`<DEL>`記号ALTアリルを用いて説明することができます。ALTアリル記号を使用して記述されたバリアントについては、`INFO/END`の値も提供する必要があります。`<INV>`ALTアリルを用いて1つのVCFレコードで表される逆位は受け入れられませんが、ブレークエンドレコードのセットに変換すれば、逆位の遺伝型を決定できます。各ブレークポイントは、ブレークエンドVCFレコードのペアによって記述されます。強制ジェノタイピングのインプットがペアの1レコードのみを含み、上記のインプット条件が満たされた場合、インプットは依然として強制ジェノタイピングのために受け入れられ、遠位ブレークエンドはローカルレコードから推定されます。

非挿入SVアリルに対するブレークポイント挿入は、次の2つのメソッドのいずれかを用いて記述することができます。どちらのメソッドも、SV VCF出力でブレークポイント挿入を記述するために使用されるフォーマットに対応しています。

- <DEL>のような記号ALTフォーマットを使用して記述されたSVの場合、INFO/SVINSSEQフィールドは、ブレークポイント挿入シーケンスを読み取るために解析されます。
- REFフィールドおよびALTフィールドに直接記載されている小さなindelについては、ALTフィールドの内容にブレークエンドシーケンスを記載します。

強制ジェノタイピングの出力

強制ジェノタイピングSVは、強制ジェノタイピングがスタンドアロンであるか、SVコールと統合されているかにかかわらず、常にSV Callerの標準VCF出力に出力されます。同じSVアリルがサンプルデータから独立して発見された場合、発見されたSVのみが最終的な出力に現れます。発見されたSVアリルは、強制ジェノタイピングのインプットSVとのマッチを示すためにアノテーションが付けられ、スコアリングおよびフィルタリングルールは、マッチするように変更されます。

強制ジェノタイピングの影響を受けたVCF出力レコードには、以下の関連フィールドがあります。

- フラグINFO/NotDiscoveredは、サンプルデータから独立して発見されなかったVCFレコードに対して設定されます。強制ジェノタイピングがスタンドアロンで実行される場合、すべての出力レコードにフラグが含まれます。SVコーリングと統合されると、フラグは標準的なSV解析では発見されなかったであろうSVアリルを区別することができます。
  - これらのバリアントの場合のみ、SV座位グラフから生成される通常のSV Caller IDフィールドは使用できません。代わりに、IDは対応するユーザーインプットVCFから取得されます。接尾辞UserInput${InputVCFRecordNumber}が、アンダースコアで区切られてIDに追加されます。インプットVCFに、ブレークエンドバリアントを構成する2つのVCFレコードのうち1つだけが含まれている場合は、メイトブレークエンドレコードから識別子が取得され、接尾辞_Mateが追加されます。
- 強制ジェノタイピングのインプットVCFレコードに対応する出力VCFレコードには、インプットVCFレコードのVCF ID値を反映するように設定された値INFO/UserInputId=${ID}がつきます。対応するレコードは、サンプルデータから独立して発見され、INFO/NotDiscoveredフラグが設定されていない場合もあります。
- インプットSVに正確にマッチする強制ジェノタイピングアリルを含む強制ジェノタイピングインプットVCFレコードに対応する任意の出力VCFレコードには、フラグINFO/KnownSVScoringがつきます。このフラグを持つVCFレコードは、常にSV Callerの出力に放出されます。MaxDepthなどのいくつかのフィルターは適用されません。

インプット要件

SV Callerを実行する場合、インプットシーケンスリードは、FRリードペア配向を有する標準的なIllumina paired-end sequencing assayからである必要があり、ここでは、各シーケンス断片について、リードは断片の各末端から内側に進みます。詳細については、189 ページの 「DRAGEN SV Callerの機能」 を参照してください。

SV Callerは、断片サイズが通常両方のリードサイズよりも大きいペアエンドライブラリー用に最適化されています。重複するリードペアを使用してSVを検出できますが、常に最適に処理されるとは限りません。典型的な断片サイズがリード長未満であるライブラリーの場合、SV Callerはアダプターシーケンスへのリードシーケンシングをバリアントシグナルと区別しようとします。このような場合、SV Callerのインプットクオリティチェックが失敗し、SV解析がスキップされる可能性があります。

スタンドアロンモードを使用する場合は、最初にBAM/CRAMインプットをマッピングする必要があります。データのマッピングとアライメントをまだ行っていない場合は、アライメントファイルを生成できます。

インプットクオリティチェック

SV Callerは、断片サイズ分布を推定する前に、各サンプルのインプットシーケンスリードのクオリティチェックを実行して、インプットが期待されるFR配向を有するペアリードアッセイに相当することを確認します。共通リードペア配向をチェックするために、高クオリティリードペアのサブセットがサンプリングされます。SV解析を続行するには、これらの少なくとも90%が予想されるFR方向を持っている必要があります。そうでない場合、SV Callerは警告を発し、それ以上の解析をスキップし、結果の出力ファイルには空の結果が表示されます。

SV Callerは、断片サイズ分布を推定するのに十分なペアエンドリードが存在する場合、インプット中の非ペアリードを許容することができます。断片サイズ分布を推定するために、SV Callerは、推定ルーチンのクオリティ要件を満たす少なくとも100個のリードペアを必要とします。ペアの両方のリードは、同じ染色体へのマッピングクオリティが0以外でなければならず、フィルタリングもスプリットリードマッピングの一部も行われず、indelもソフトクリッピングも含みません。サンプルにこのようなリードペアが十分な数含まれていない場合、SV Callerは警告を発し、それ以上の解析をスキップして、空の結果を出力ファイルに書き込みます。

リードグループ

SV Callerは、インプットシーケンスに適用されたリードグループラベルをすべて無視します。各インプットサンプルは、単一の断片サイズ分布を有する別個のライブラリーとして処理されます。

ファイルフォーマット

スタンドアロンモードでは、インプットシーケンスリードはマッピングされ、BAMまたはCRAMフォーマットでインプットとして提供される必要があります。各入力ファイルは座標ソートされ、インデックスが作成されて、インプットBAMまたはCRAMファイルとマッチする名前のファイルにhtslibスタイルのインデックスが作成されます。ファイル名の拡張子には`*.bai`、`*.cali`、または`*.csi`が追加されます。スタンドアロンモードの詳細については、190 ページの 「操作モード」 を参照してください。

正常サンプルまたは腫瘍サンプルについて、少なくとも1つのBAMファイルまたはCRAMファイルを提供する必要があります。マッチするTumor-Normalサンプルペアも同様に提供することができます。複数の入力ファイルが正常サンプルに提供される場合、各ファイルは、ジョイント二倍体サンプル解析の一部として別個のサンプルとして扱われます。

スタンドアロンモードでは、インプットBAMまたはCRAMファイルに次の制限があります。

- アライメントに不明なリードシーケンスを含めることはできません（SEQ=「*」）
- アライメントのSEQフィールドに「=」文字を含めることはできません。

- アライメントでは、シーケンスのマッチ/アンマッチ（「=」/「X」）は使用できません。アライメントレコード内のCIGAR表記RG（リードグループ）タグは無視されます。各アライメントファイルは、1つのサンプルを表すものとして扱われます。
- ベースコールクオリティ値が70を超えるアライメントは拒否されます。これらは、オフセットエラーを示すという前提ではサポートされていません。

アライメントファイルの生成

以下のコマンドラインの例では、入力タイプに応じてDRAGENのマップ/アライメントパイプラインを実行する方法について示します。マップ/アライメントパイプラインでは、パイプラインで使用できるBAMまたはCRAMファイルの形式でアライメントファイルを生成します。

まだマップもアライメントもされていないすべてのサンプルのアライメントファイルを生成する必要があります。各サンプルには、ユニークなサンプル識別子が備えられている必要があります。*--RGSM*オプションを使用して、識別子を指定します。BAMおよびCRAM入力ファイルでは、サンプル識別子がファイルから取得されるため、*--RGSM*オプションは必要ありません。

以下は、コマンドがFASTQファイルをマップしてアライメントする例です：

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

以下は、コマンドが既存のBAMファイルをマップしてアライメントする例です：

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

以下は、コマンドが既存のCRAMファイルをマップしてアライメントする例です：

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

エクソーム/ターゲットコール

SV Callerは、ターゲットシーケンスインプット用に設定できます。これにより、高深度フィルターが無効になります。エクソームモードは、コマンドラインオプション*--sv-exome*を使用して、trueまたはfalseに直接設定できます。直接設定しない場合、エクソームモードは初期設定でfalseに設定されます。ただし、SV Callerを統合モードで実行し、シーケンスインプットが50 Gb以下である場合は除きます。

内部タンデム重複のターゲットコール

内部タンデム重複（ITD）を対象とした解析では、`--sv-use-overlap-pair-evidence true`コマンドラインオプションを使用して、重複するリードペアをSVエビデンスと見なすことで、SV Callerの感度を高めることができます。設定されていない場合、重複するリードペアの割合が20%未満になると、コマンドは自動的に有効になります。

`--sv-somatic-ins-tandup-hotspot-regions-bed${BEDFILE}`オプションを使用すると、ITDホットスポット領域を指定して、体細胞バリアント解析におけるITDコールの感度を上げることができます。初期設定では、DRAGEN SVは`/opt/edico/config/sv_somatic_ins_tandup_hotspot_*.bed`からリファレンス固有のホットスポットBEDファイルを自動的に選択します。このファイルには、FLT3、ARHGEF7、およびKMT2Aが含まれます。この機能を無効にするには、`--sv-enable-somatic-ins-tandup-hotspot-regions false`を入力します。

血液腫瘍コール

血液腫瘍モードの血液腫瘍は、白血病のような血液がんのことを言います。Tumor-Normal解析では、DRAGENは血液腫瘍モードを実行することにより、Tumor-in-Normal（TiN）コンタミネーションを考慮します。体細胞状態の事後確率を計算するときに、マッチする正常についてバリアントアリル頻度がゼロでないことを許容することによって、血液腫瘍モードを使用してTiNコンタミネーションを考慮できます。コンタミネーションが考慮されない場合、真の体細胞バリアントを抑制することにより、感度に重大な影響を与える場合があります。

血液腫瘍モードの動作を制御するには、次の2つのオプションを使用します。

- `--sv-enable-liquid-tumor-mode`：血液腫瘍モードを有効にします。血液腫瘍モードは、初期設定で無効にされています。
- `--sv-tin-contam-tolerance`：TiNコンタミネーションの許容レベルを設定します。DRAGENは、指定された最大許容レベルまでのTiNコンタミネーションの存在下でバリアントをコールします。0～1からまでの任意の値を入力できます。初期設定の最大TiNコンタミネーション許容値は0.15です。初期設定値を用いる場合、体細胞バリアントは、腫瘍サンプル中の対応するアリルの15%までのアリル頻度で正常サンプル中に検出されると予想されます。

コマンドラインオプション

構造多型コーラーでは、次のコマンドラインオプションがサポートされています。

インプットおよびアウトプットオプション

DRAGEN SVパイプラインは、DRAGENホストソフトウェアと以下のインプットおよびアウトプットオプションを共有します。BAMファイルとCRAMファイルをインプットとして使用できます。あるいは、リードマッピングとSVコールを1回の実行で使用する場合は、FASTQ、BAM、CRAMファイルなど、DRAGENインプットオプションをすべて使用できます。

オプション	説明
--bam-input	解析対象のBAMファイル。
--tumor-bam-input	Tumor-Normal解析またはTumor-only解析を行う場合の解析対象BAMファイルをインプットします。
--cram-input	解析対象のCRAMファイル。

オプション	説明
--tumor-cram-input	Tumor-Normal解析またはTumor-only解析を行う場合の解析対象CRAMファイルをインプットします。
--enable-map-align	DRAGENマップ/アライメントを有効にします。初期設定はtrueであるため、このオプションをfalseに設定しない限り、すべての入力リードが再マッピングされてアライメントされます。
--fastq-file1, --fastq-file2, --fastq-list	FASTQファイルまたは解析対象ファイルのリストをインプットします。
--tumor-fastq1, --tumor-fastq2, --tumor-fastq-list	腫瘍FASTQファイルまたは解析対象ファイルのリストをインプットします。
--ref-dir	DRAGENリファレンスゲノムのハッシュテーブルディレクトリ。リファレンスゲノムのハッシュテーブルについて詳しくは、10 ページの「リファレンスゲノムの準備」を参照してください。
--output-directory	すべての結果が格納される出力ディレクトリ。
--output-file-prefix	すべての結果ファイル名の前に付加される出力ファイルの接頭辞。

SVコーリングオプション

オプション	説明
--enable-sv	構造多型コーラーを有効または無効にします。初期設定はfalseです。
--sv-call-regions-bed	コールする領域のセットを含むBEDファイルを指定します。オプションで、ファイルをgzip形式またはbgzip形式で圧縮できます。
--sv-region	デバッグのためにゲノムの特定の領域に解析を限定します。このオプションを複数回指定すると、領域のリストを作成できます。値の形式は、「chr:startPos-endPos」である必要があります。
--sv-exome	trueに設定すると、ターゲットシーケンスインプットのバリアントコーラーが設定されます。これには、高深度フィルターの無効化も含まれます。統合モードでは、初期設定でターゲットシーケンスインプットを自動検出します。スタンドアロンモードでは、初期設定はfalseです。
--sv-output-contigs	Trueに設定すると、アセンブルされたコンティグシーケンスがVCFファイルに出力されます。初期設定はfalseです。
--sv-forcegt-vcf	強制ジェノタイピングのための構造多型のVCFを特定します。バリアントは、サンプルデータで見つからない場合でも、スコアリングされて出力VCFに放出されます。バリアントは、サンプルデータから直接発見された追加のバリアントとマージされます。

オプション	説明
--sv-discovery	SV発見を有効にします。このフラグは、--svforcegt-vcfが使用されている場合にのみfalseに設定できます。falseに設定すると、SV発見は無効になり、強制ジェノタイピングのインプットバリアントのみが処理されます。初期設定はtrueです。
--sv-use-overlap-pair-evidence	重複するリードペアをエビデンスと見なすことを許可します。初期設定では、DRAGENは重複するリードペアの割合が20%未満の場合に自動検出を使用します。
--sv-somatic-instandup-hotspot-regions-bed	ITDホットスポット領域のBEDを指定して、体細胞バリアント解析におけるITDコールの感度を高めます。初期設定では、DRAGEN SVは`/opt/edico/config/sv_somatic_ins_tandup_hotspot_*.bed`からリファレンス固有のホットスポットBEDファイルを自動的に選択します。
--sv-enable-somatic-ins-tandup-hotspot-regions	ITDホットスポット領域のインプットを有効または無効にします。体細胞バリアント解析では初期設定はtrueです。
--sv-enable-liquid-tumor-mode	血液腫瘍モードを有効にします。詳細については、198 ページの「血液腫瘍コール」を参照してください。
--sv-tin-contam-tolerance	Tumor-in-Normal（TiN）コンタミネーション許容レベルを設定します。詳細については、198 ページの「血液腫瘍コール」を参照してください。

構造多型VCF出力

構造多型VCF出力ファイルは、出力ディレクトリにあります。ファイル名は`<output-file-prefix>.sv.vcf.gz`です。ファイルの内容は、解析のタイプによって異なります。

主要な解析カテゴリー（生殖細胞系列、Tumor-Normal、Tumor-only）ごとに、適切なVCF出力ファイルが出力され、指定した解析タイプに対応するバリアントコールモードの下で行われたバリアントコールが反映されます。

構造多型コーラーは`<output-directory>/sv/`ディレクトリに追加の出力を生成します。`<output-directory>/sv/results`サブディレクトリには、追加のバリアントと統計出力ファイルが含まれます。追加のサブディレクトリには、バリアントコールプロセスからのログと中間出力が含まれます。

構造多型の予測

`<output-directory>/sv/results/variants`では、SV Callerは一連のVCFファイルを出力します。生殖細胞系列解析用に2つのVCFファイルが作成されます。Tumor-Normal解析では、Tumor-Normalを差し引くために追加の体細胞VCFを作成します。Tumor-only解析では、SV Callerは追加の出力VCFを生成します。次のVCFファイルが出力されます：

ファイル	説明
`diploidSV.vcf.gz`	ジョイント二倍体サンプル解析ではサンプルセットについて、Tumor-Normal減法解析では正常サンプルについて二倍体モデルを用いてスコアリングとジェノタイピングを行われたSVとindel。Tumor-Normal減法解析の場合、このファイルのスコアは腫瘍サンプルからの情報を反映していません。
`somaticSV.vcf.gz`	体細胞バリアントモデルでもスコアリングされたSVとindel。このファイルは、設定中に腫瘍サンプルアライメントファイルが提供された場合にのみ作成されます。
`candidateSV.vcf.gz`	スコアリングされていないSVおよびindel候補。バリアントがこのファイルに候補としてインプットされるためには、最小限の裏付け証拠のみが必要です。バリアントは、スコアリングの対象となる候補である必要があります。したがって、バリアントがファイルに存在しない場合、他のVCF出力には表示されません。このファイルには、サイズ8以上のindelが含まれています。ワークフローを柔軟にするために最小のindelが提供されますが、スコアリングは行われません。Indelスコアリングはサイズ50から始まります。
`tumorSV.vcf.gz`	重複する候補および最小スコア付きバリアントサイズ(50)未満の小さなindelを削除した後の`candidateSV.vcf.gz`ファイルのサブセット。SVはスコアリングされませんが、次の詳細情報が追加されます: • 各アリルのペアリードおよびスプリットリード裏付け証拠カウント • スコアリングしたTumor-Normalモデルからのフィルターのサブセットは、精度を改善するために単一の腫瘍ケースに適用されます。

VCF出力

VCF出力は、構造多型を記述するためのVCF 4.1仕様に準拠しています。可能な限り標準のフィールド名を使用します。カスタムフィールドはすべて、VCFヘッダーに記述されます。次のセクションでは、バリアントの詳細とプライマリーVCFフィールド値について説明します。

VCFサンプル名

VCF出力に出力されるサンプル名は、ヘッダーにある最初のリードグループ（@RG）レコードからの各インプットアライメントファイルから抽出されます。サンプル名が見つからない場合は、代わりに初期設定のラベル（SAMPLE1、SAMPLE2など）が使用されます。

スモールindel

すべてのバリアントは、スモールindelに分類されない限り、記号アリルを用いてVCFでレポートされ、その場合、VCF REFおよびALTアリルフィールドの完全なシーケンスが提供されます。バリアントは、以下のすべての基準を満たす場合、スモールindelと分類されます：

- バリアントは、挿入されたシーケンスと欠失したシーケンスの組み合わせとして完全に表現することができます。
- 欠失または挿入の長さが1000以上ではありません。
- バリアントのブレークエンドおよび/または挿入されたシーケンスは不精確ではありません。
- このバリアントは、深度に基づくSV分類ルーチンによって欠失から染色体内ブレークエンドに変換されていません。

VCFレコードがスモールindelフォーマットで出力される場合、挿入イベントと欠失イベントの組み合わせを記述するCIGARINFOタグも含まれます。

不完全なシーケンスアセンブリを伴う挿入

挿入シーケンスが完全にアセンブルされない場合でも、大きな挿入がレポートされることがあります。この場合、SV Callerは<INS>記号アリルを使用して挿入をレポートし、特殊なINFOフィールドLEFT_SVINSSEQとRIGHT_SVINSSEQを含めて、挿入シーケンスのアセンブルされた左端と右端を記述します。以下は、hg19にマッピングされたNA12878、NA12891およびNA12892のジョイント二倍体解析から得られたそのようなレコードの例です：

```
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=TAAATTTTTC
TT;LEFT_
SVINSSEQ=TAAATTTTTCTTTTTTCTTTTTTTTTTAAATTTATTTTTTTATTGATAATTCTTGGGTGTTTCT
CACAGAGGGGGATTTGGCAGGGTCACGGGACAACAGTGGAGGGAAGGTCAGCAGACAAACAAGTGAACAAAGG
TCTCTGGTTTTCCCAGGCAGAGGACCCTGCGGCCTTCCGCAGTGTTCGTGTCCCTGATTACCTGAGATTAGGG
ATTTGTGATGACTCCCAACGAGCATGCTGCCTTCAAGCATCTGTTCAACAAAGCACATCTTGCACTGCCCTTA
ATTCATTTAACCCCGAGTGGACACAGCACATGTTTCAAAGAG;RIGHT_
SVINSSEQ=GGGGCAGAGGCGCTCCCCACATCTCAGATGATGGGCGGCCAGGCAGAGACGCTCCTCACTTCCT
AGATGTGATGGCGGCTGGGAAGAGGCGCTCCTCACTTCCTAGATGGGACGGCGGCCGGGCGGAGACGCTCCTC
ACTTTCCAGACTGGGCAGCCAGGCAGAGGGGCTCCTCACATCCCAGACGATGGGCGGCCAGGCAGAGACACTC
CCCACTTCCCAGACGGGGTGGCGGCCGGGCAGAGGCTGCAATCTCGGCACTTTGGGAGGCCAAGGCAGGCGGC
TGCTCCTTGCCCTCGGGCCCCGCGGGGCCCGTCCGCTCCTCCAGCCGCTGCCTCC 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
```

小さなタンデム重複のノーマライズ

SV Callerはタンデム重複を挿入として表現することもできます。この表現は、特に小さなタンデム重複の場合に、VCF出力におけるバリアントの提示方法に曖昧さを生じさせます。この表現は、認識されないコール重複などの複雑な問題を引き起こす可能性があります。

同じバリアントが2つの異なるVCFフォーマットで表現されないように、SV Callerの出力をより正確にノーマライズするために、小さなタンデム重複（1,000塩基未満）をVCF出力の挿入に変換します。このようなタンデム重複から変換された挿入は、不完全挿入に似たフォーマットをもち、ALTフィールドには記号アリル<INS>を用います。次の例は、このノーマライゼーションプロセス中にタンデム重複から変換された挿入を示しています。

```
chr2 2520057 MantaDUP:TANDEM:53645:0:1:0:0:0 T <INS> 813 PASS
END=2520057;SVTYPE=INS;SVLEN=52;DUPSVLEN=52 GT:FT:GQ:PL:PR:SR
0/1:PASS:393:863,0,390:25,0:19,25
```

変換された挿入には、特定の出力フィールドのコピーが含まれます。このフィールドは、タンデム重複レコードの場合と同じように表示されます。例えば、INFO/DUPSVINSSEQは、重複に対して計算されたブレークポイント挿入値のコピーを提供します。重複のコンテキストでは、ブレークポイント挿入値は通常、INFO/SVINSSEQに書き込まれます。次の例は、ブレークポイント挿入値を持つ変換された挿入を示しています。

```
chr2 2645730 MantaDUP:TANDEM:53649:0:1:0:0:0 C <INS> 367 PASS
END=2645730;SVTYPE=INS;SVLEN=97;DUPSVLEN=86;DUPSVINSLEN=11;DUPSVINSSEQ=CT
CACCTTCAT GT:FT:GQ:PL:PR:SR 0/1:PASS:367:417,0,386:19,0:20,15
```

コピーされたINFOフィールドの詳細については、207 ページの 「VCF INFOフィールド」 を参照してください。すべてのINFOフィールドで同じDUP接頭辞が使用されます。

逆位

逆位は一連のブレークエンドとしてレポートされます。例えば、相互の逆位の場合、同じEVENT INFOタグを共有する4つのブレークエンドがレポートされます。次に、相互の逆位を表すブレークエンドレコードの例を示します：

```
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
```

深度ベースのSVタイプ分類

生殖細胞系列コーリングモデルでは、サンプルデータからSV候補が発見され、出力レポートに十分なペアリードおよびスプリットリードの証拠がある場合、SV Callerは追加の深度ベースの検証を行って、特定のSV候補タイプをより正確に分類します。欠失と一致する候補ブレークポイントは、欠失領域内で予想されるより低いリード深度について検証されます。タンデム重複と一致する候補ブレークポイントは、重複領域で予想されるより高いリード深度について検証されます。深度ベースの検証に失敗した候補SVコールは出力時にレポートされますが、染色体内ブレークエンドに変更されます。通過する候補SVコールは、標準の欠失およびタンデム重複の出力フォーマットで引き続きレポートされます。

SVブレークポイント

SVブレークポイントの挿入

SVはしばしばブレークポイントに小さなシーケンスの挿入を含みます。ブレークポイントの挿入は、SVのタイプによって表示が異なります。VCF出力の`INFO/SVINSSEQ`フィールドは、挿入シーケンスそのものを記述することによって、ブレークポイント挿入の最も一般的な記述を提供します。対応する`INFO/SVINSLEN`フィールドは、挿入シーケンスの長さを記述します。例えば、次のVCFレコードは大きな（約8.8 kb）欠失を記述しており、これには左右の欠失ブレークエンドの間に1塩基挿入（`C`）が含まれています。

```
chr22 17770350 MantaDEL:101:0:1:0:0:0 C <DEL> 687 PASS
END=17779108;SVTYPE=DEL;SVLEN=-8758;SVINSLEN=1;SVINSSEQ=C
GT:FT:GQ:PL:PR:SR 0/1:PASS:687:737,0,858:39,20:32,8
```

`INFO/SVINSSEQ`フィールドは、タンデム重複およびブレークエンドレコードのブレークポイント挿入を記述するためにも使用されます。このフィールドは、大きなSV挿入の挿入シーケンスを記述するためにも使用できます。

VCFのスモールindelフォーマットでは、ブレークポイントの挿入は異なって表されます。SV Callerは、記号ALTアリルの代わりにVCFのスモールindelフォーマットを用いて、スモール欠失および挿入を表します。VCFのスモールindelフォーマットに生じるブレークポイント挿入はすべて、VCF ALTフィールドの一部として表されます。このフォーマットがSVに使用される条件については、202 ページの 「スモールindel」を参照してください。

以下のスモールindelフォーマットの例では、VCFレコードは、左右の欠失ブレークエンド間の1塩基挿入（A）を含む57塩基欠失を記述します。

```
chr22 32981929 MantaDEL:1136:0:0:0:0:0
TGTATACATATATGTGTATATACGTATATATGTATATATGTATGTATACGTATATATG TA 537 PASS
END=32981986;SVTYPE=DEL;SVLEN=-57;CIGAR=1M1I57D GT:FT:GQ:PL:PR:SR
0/1:PASS:308:587,0,305:8,0:23,15
```

ブレークエンドレコードには、ブレークエンドALTフィールドのVCF仕様に記述されているとおり、ブレークポイント挿入シーケンスの追加コード化が含まれています。SV Callerは、他のSVレコードタイプとの一貫性を保つために、INFO/SVINSSEQフィールドにも情報を提供します。

次の例は、サンプル中の第1染色体と第12染色体の領域を、2つのブレークエンドの間のCAのブレークエンド挿入シーケンスに結合するブレークエンドを示します。挿入シーケンスは、ALTフィールドとINFO/SVINNSEQフィールドの両方に記述されます。

```
1 39604587 MantaBND:31780:1:3:0:0:0:1 T TCA[12:6472102[ 774 PASS
SVTYPE=BND;MATEID=MantaBND:31780:1:3:0:0:0:0;SVINSLEN=2;SVINSSEQ=CA;BND_
DEPTH=67;MATE_BND_DEPTH=55 GT:FT:GQ:PL:PR:SR
0/1:PASS:774:824,0,999:63,3:36,33
12 6472102 MantaBND:31780:1:3:0:0:0:0 G ]1:39604587]CAG 774 PASS
SVTYPE=BND;MATEID=MantaBND:31780:1:3:0:0:0:1;SVINSLEN=2;SVINSSEQ=CA;BND_
DEPTH=55;MATE_BND_DEPTH=67 GT:FT:GQ:PL:PR:SR
0/1:PASS:774:824,0,999:63,3:36,33
```

SVブレークポイントの挿入配向

ブレークポイント挿入シーケンスは、常に現在のSVレコードのストランドに関して提供されます。ブレークエンドレコードの中には、配向が反転しているものがあります。反転配向の場合、ブレークエンドレコードのペアには、メイトされたレコードと比較して逆相補した挿入シーケンスが含まれます。

次のブレークエンドペアの例は、反転配向を示しています。

```
1 210891730 MantaBND:43882:0:2:0:2:0:1 A AATG]19:45732595] 999 PASS
SVTYPE=BND;MATEID=MantaBND:43882:0:2:0:2:0:0;SVINSLEN=3;SVINSSEQ=ATG;BND_
DEPTH=76;MATE_BND_DEPTH=106 GT:FT:GQ:PL:PR:SR
0/1:PASS:999:999,0,999:69,16:43,55
19 45732595 MantaBND:43882:0:2:0:2:0:0 G GCAT]1:210891730] 999 PASS
SVTYPE=BND;MATEID=MantaBND:43882:0:2:0:2:0:1;SVINSLEN=3;SVINSSEQ=CAT;BND_
DEPTH=106;MATE_BND_DEPTH=76 GT:FT:GQ:PL:PR:SR
0/1:PASS:999:999,0,999:69,16:43,55
```

SVブレークポイントの相同性

SV Callerによって出力された各VCFレコードは、ブレークポイントの正確な相同性範囲の左端にシフトされます。ブレークポイントの正確な相同性範囲は、同じSVハプロタイプを示しながらSVを示す位置の連続的な範囲です。正確な相同性範囲は、正確な相同性範囲のシーケンスを記述するINFO/HOMSEQフィールドと、それに対応し、範囲の長さを記述するINFO/HOMLENフィールドとともにVCF出力に記述されます。

以下の実施例は、11塩基のブレークエンド相同領域を有する62塩基欠失を示します。左シフトしない場合、SVは位置39497639から39497650までの任意の場所で同様に示されます。

```
chr22 39497639 MantaDEL:34:85:85:1:0:0
GGGGGGTGGGGGCGGGTTGGAGGAGGTTGGCGGGGGGCGGGGGCGGGTTGGAGGAGGTTGGCA G 187
PASS END=39497701;SVTYPE=DEL;SVLEN=-
62;CIGAR=1M62D;CIPOS=0,11;HOMLEN=11;HOMSEQ=GGGGGTGGGGG GT:FT:GQ:PL:PR:SR
0/1:PASS:12:237,0,8:4,0:2,8
```

以下の例は、単純化した正確なブレークエンドの相同性を示します。この例では、1つの3塩基欠失ともう1つの3塩基挿入が示されています。挿入と欠失の両方で、バリアントは左シフトされるため、対応するVCFレコードの位置は2になります。

欠失

```
Reference: GTCAGCGA

Variant: GT---CGA
```

挿入

```
Reference: GT---CAG

Variant: GTCGGCAA
```

挿入でも欠失でも、ブレークエンド相同性Cの塩基は1つしかないため、同じバリアントを1塩基右に示します。

VCF INFOフィールド

ID	説明
IMPRECISE	構造バリエーションが不精確であること、すなわち正確なブレークポイントの位置が見つからないことを示すフラグ
SVTYPE	構造多型のタイプ
SVLEN	REFアリルとALTアリル間の長さの差
END	このレコードに記述されているバリアントの終了位置
CIPOS	POS付近の信頼区間
CIEND	END付近の信頼区間
CIGAR	各代替indelアリルのCIGARアライメント
MATEID	メイトブレークエンドの識別番号
EVENT	ブレークエンドに関連付けられたイベントのID
HOMLEN	イベントのブレークポイントにおける塩基対の同一相同性の長さ
HOMSEQ	イベントのブレークポイントにおける塩基対の同一相同性のシーケンス
SVINSLEN	挿入の長さ
SVINSSEQ	挿入のシーケンス
LEFT_SVINSSEQ	未知の長さの挿入のための既知の挿入左側
RIGHT_SVINSSEQ	未知の長さの挿入に対する既知の挿入右側
PAIR_COUNT	このバリアントをサポートするリードペア。ここで、両方のリードは確信を持ってマッピングされます
BND_PAIR_COUNT	このブレークエンドにおいて、このバリアントを裏付ける、確信を持ってマッピングされたリード(リモートブレークエンドではマッピングが信頼できない可能性がある)
UPSTREAM_PAIR_COUNT	上流のブレークエンドにおいて、このバリアントを裏付ける、確信を持ってマッピングされたリード(下流のブレークエンドではマッピングが信頼できない可能性がある)
DOWNSTREAM_PAIR_COUNT	下流のブレークエンドにおいて、このバリアントを裏付ける、確信をもってマッピングされたリード(上流のブレークエンドではマッピングが信頼できない可能性がある)
BND_DEPTH	ローカル転座ブレークエンドでのリード深度
MATE_BND_DEPTH	リモート転座メイトブレークエンドでのリード深度
JUNCTION_QUAL	SVジャンクションがEVENTの一部である場合(すなわち、複数隣接バリアント)、このフィールドは、問題の隣接のみのQUAL値を提供します。
SOMATIC	体細胞バリアントを示すフラグ
SOMATICSCORE	体細胞バリアントのクオリティスコア

ID	説明
JUNCTION_SOMATICSCORE	SVジャンクションがイベントの一部である場合（すなわち、複数隣接バリアント）、このフィールドは、問題の隣接のSOMATICSCORE値だけを提供します。
CONTIG	アセンブルされたコンティグシーケンス（バリアントが不精確でない場合）（*--outputContig*）
DUPSVLEN	重複したリファレンスシーケンスの長さ
DUPHOMLEN	重複したリファレンスシーケンスを除くイベントブレークポイントにおける塩基対の同一相同性の長さ
DUPHOMSEQ	重複したリファレンスシーケンスを除くイベントブレークポイントにおける塩基対の同一相同性のシーケンス
DUPSVINSLEN	重複したリファレンスシーケンス後の挿入されたシーケンスの長さ
DUPSVINSSEQ	重複したリファレンスシーケンス後の挿入されたシーケンス
NotDiscovered	ユーザーによって指定された、インプットシーケンスデータでは発見されなかったバリアント候補
UserInputId	ユーザーインプットVCFからのバリアントID
KnownSVScoring	バリアントは、ユーザー指定のインプットバリアントに関連付けられています。そのため、スコアリングとフィルタリングの基準は、より強力な以前の真の仮定の下で緩和されます。

VCF FORMATフィールド

ID	説明
GT	遺伝型
FT	サンプルフィルター。`PASS`は、すべてのフィルターがこのサンプルに合格したことを示します。
GQ	遺伝型クオリティ
PL	VCF仕様で定義されている遺伝型のノーマライズされたPhred値の尤度
PR	REFアリルまたはALTアリルを強く（Q30）裏付けるスパンニングリードペアの数
SR	REFアリルまたはALTアリルを強く（Q30）裏付けるスプリットリードの数

VCF FILTERフィールド

生殖細胞系列

次の表に、生殖細胞系列VCF出力に適用されるVCF FILTERフィールドを示します。

ID	レベル	説明
MinQUAL	レコード	QUALスコアは20未満。このフィルターは、KnownSVScoringフラグが設定されているレコードには適用されません。
MinGQ	サンプル	GQスコアは15未満。このフィルターはサンプルレベルで適用され、KnownSVScoringフラグの付いたレコードには適用されません。
Ploidy	レコード	DELバリアントとDUPバリアントの場合、類似したサイズの重複バリアントの遺伝型は、二倍体の期待値と一致しません。このフィルターは、KnownSVScoringフラグが設定されているレコードには適用されません。
MaxDepth	レコード	深度は、1つまたは両方のバリアントブレークエンド付近の染色体の深度の中央値の3倍を超えます。このフィルターは、KnownSVScoringフラグが設定されているレコードには適用されません。
MaxMQ0Frac	レコード	スモールバリアント（1000塩基未満）の場合、いずれかのブレークエンド周辺でMAPQ0を有する全サンプル中のリードの割合は0.4を超えます。このフィルターは、KnownSVScoringフラグが設定されているレコードには適用されません。
NoPairSupport	レコード	ペアリード断片サイズよりも大幅に大きいバリアントの場合、どのサンプルにおいても、代替アリルを裏付けるペアリードは存在しません。このフィルターは、KnownSVScoringフラグが設定されているレコードには適用されません。
SampleFT	レコード	すべてのサンプルレベルフィルターを通過するサンプルはありません。
HomRef	サンプル	ホモ接合性リファレンスコール。このフィルターはサンプルレベルで適用されます。

Tumor-Normal体細胞

下表は、Tumor-Normal体細胞VCF出力に適用されるVCF FILTERフィールドをリストしたものです。

ID	レベル	説明
MinSomaticScore	レコード	SOMATICSCOREは30未満。
MaxDepth	レコード	正常なサンプル部位の深度は、一方または両方のバリアントブレークエンド付近の染色体の深度の中央値の3倍を超えます。このフィルターは、KnownSVScoringフラグが設定されているレコードには適用されません。
MaxMQ0Frac	レコード	正常サンプル中のスモールバリアント（1000塩基未満）では、いずれかのブレークエンド周辺のMAPQ0のリードの割合は0.4を超えます。このフィルターは、KnownSVScoringフラグが設定されているレコードには適用されません。

Tumor-only

下表に、Tumor-onlyのVCF出力に適用されるVCF FILTERフィールドを示します。

ID	レベル	説明
MaxDepth	レコード	正常なサンプル部位の深度は、一方または両方のバリアントブレークエンド付近の染色体の深度の中央値の3倍を超えます。このフィルターは、KnownSVScoringフラグが設定されているレコードには適用されません。
MaxMQ0Frac	レコード	スモールバリアント（1000塩基未満）の場合、いずれかのブレークエンド周辺のMAPQ0のリードの割合は0.4を超えています。このフィルターは、KnownSVScoringフラグが設定されているレコードには適用されません。

VCFフィルターの解釈

VCFフィルターには、レコードレベル（FILTER）とサンプルレベル（FORMAT/FT）の2つのレベルがあります。

ほとんどのレコードレベルフィルターは、サンプルレベルフィルターとは独立しています。ただし、生殖細胞系列解析では、すべてのサンプルレベルフィルターを通過するサンプルがない場合は、SampleFTフィルターレコードレベルフィルターが適用されます。

INFO/EVENTフィールドの解釈

転座など、VCFでレポートされている構造多型の一部は、サンプル中の単一の新規シーケンスジャンクションを表します。INFO/EVENTフィールドは、2つ以上のそのようなジャンクションが単一のバリアントイベントの一部として一緒に発生すると仮定されることを示します。同一イベントに属する個々のバリアントレコードはすべて、同じINFO/EVENT文字列を共有します。このような推定は、コールされたバリアントブレークポイントの相対的距離と配向を解析することによってSVコールの後に適用することができますが、SV Callerはこのイベントメカニズムをコールプロセスに組み込んで、このような大規模イベントに対する感度を高めることにご留意ください。イベント中の少なくとも1つのジャンクションがすでに標準的なバリアント候補の閾値を超えていると仮定すると、マルチジャンクションイベント（相互転座対など）と一致するパターンで生じる付加的なジャンクションの証拠閾値を下げることによって感度が改善されます。

このメカニズムは、任意の数のジャンクションを含むイベントに一般化することができますが、現在は2つに限られています。そのため、現時点では相互転座対を同定し、感度を改善するために最も有用です。

VCF IDフィールド

VCF IDまたは識別子フィールドはアノテーションに使用することができます。また、転座に関するBND（ブレークエンド）レコードの場合は、ブレークエンドのメイトまたはパートナーをリンクするためにID値を使用します。次に、SV CallerからのVCF IDフィールドの例を示します。

```
MantaINS:1577:0:0:0:3:0
```

IDフィールドで提供される値は、SVまたはindelが発見されたSV関連グラフのエッジを反映します。この値は、SV Callerによって生成された単一のVCF出力ファイル内でユニークであり、標準のVCF MATEIDキーを使用して関連するブレークエンドレコードをリンクするために使用されます。値全体をユニークキーとして使用できますが、キーを解析すると、将来のDRAGENバージョンとの互換性が失われる可能性があります。DRAGENの最新版については、DRAGEN Bio-It Platformのサポートサイトを参照してください。

SV VCFのBEDPEフォーマットへの変換

構造多型をBEDPEフォーマットで表現すると都合が良い場合があります。そのようなアプリケーションの場合、DRAGENはGitHubで利用可能なスクリプトvcfToBedpeを推奨します。リポジトリーは@hall-labから分岐されており、VCF 4.1 SVフォーマットをサポートするように変更されています。

BEDPEフォーマットでは、SV Caller VCF出力と比較して構造多型情報が大幅に削減されます。特に、座位情報やサンプル特異的な情報のためのフィールドを規定する能力に加えて、ブレークエンド配向、ブレークエンド相同性および挿入シーケンスがなくなります。この理由から、イルミナでは、必要とするアプリケーションの一時出力としてのみBEDPEを推奨します。

統計情報出力ファイル

追加の統計情報は、`<output-directory>/sv/results/stats`のファイルで提供されます。

ファイル	説明
`alignmentStatsSummary.txt`	各インプットアライメントファイルの断片長の分位。
`svLocusGraphStats.tsv`	SV座位グラフに関する統計情報およびランタイム情報。
`svCandidateGenerationStats.tsv`	SV候補生成に関する統計情報およびランタイム情報。
`svCandidateGenerationStats.xml`	svCandidateGenerationStats.tsvレポートを裏付けるXMLデータ。
`diploidSV.sv_metrics.csv`	二倍体モデルの下で通過するSVコール数。このファイルは生殖細胞系列解析またはTumor-Normal解析でのみ作成されます。
`somaticSV.sv_metrics.csv`	体細胞バリアントモデルの下で通過するSVコール数。このファイルはTumor-Normal解析でのみ作成されます。
`tumorSV.sv_metrics.csv`	Tumor-only解析で通過するSVコール数。このファイルはTumor-only解析でのみ作成されます。

構造多型のde novoクオリティスコアリング

DRAGENでは、*de novo*構造多型のクオリティスコアリングを有効化できます。

構造多型ジョイント二倍体コールの*de novo*スコアリングを有効にするには、`--sv-denovo-scoring`をtrueに設定します。バリアントが*de novo*として分類される閾値を調整するには、`--sv-denovo-threshold`コマンドラインオプションを使用します。詳細については、212 ページの 「DNフィールド」を参照してください。

インプット

*de novo*スコアリングには、次の2つのファイルが必要です：

- 系統中のサンプルすべての関係を明記したpedigreeファイル。
- 系統内のサンプルすべてに対して一緒に実行される生殖細胞系列構造多型コール解析から得たVCF出力。

pedigreeファイル

*de novo*スコアリングにはpedigreeファイルが必要です。ジョイントスモールバリアントコール解析と*de novo*スコアリングに必要とされるファイル形式と同じファイル形式を使用します。ファイルフォーマットについて詳しくは、119 ページの 「スモールバリアントDe Novoコール」 を参照してください。このファイルには、トリオの中のどのサンプルが発端者、母親、または父親であるかが明記されています。pedigreeファイルに複数のトリオが指定されている場合（例：複数世代の血統または兄弟姉妹）、DRAGENは自動的にトリオを検出し、検出された各トリオの発端者サンプルの*de novo*スコアを提供します。

ジョイント生殖細胞系列構造多型VCF

DRAGENは、pedigreeファイルに明記されたサンプルすべてについて、生殖細胞系列構造多型解析から得たVCF出力に*de novo*スコアリングを適用します。コマンドラインを用いてVCFファイルを直接提供することも、*de novo*スコアリングが有効になっているDRAGENランの一部としてファイルを生成することもできます。

出力

*de novo*スコアリングでは、出力VCFファイル内の各サンプルに対して、*de novo*クオリティスコア（*DQ*）フィールドと*de novo*コール（`DN`）フィールドが追加されます。

DQフィールド

`DQ`フィールドは、以下のように定義されます。

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

`DQ`フィールドは、発端者のバリアントが*de novo*であるというPhred値の事後確率を表します。例えば、DQスコア13および20は、*de novo*バリアントの事後確率0.95および0.99に相当します。DRAGENがDQスコアを計算できる場合は、そのスコアを発端者のサンプルに加算します。DQスコアを計算できない場合、フィールドは「.」に設定されます。

DNフィールド

`DN`フィールドは次のように定義されます。

```
##FORMAT=<ID=DN,Number=1,Type=String,Description="Possible values are
'DeNovo' or 'LowDQ'. Threshold for a passing de novo call is DQ >= 20">
```

DRAGENは有効な（>0）DQスコアを閾値と比較します。`--sv-denovo-threshold`コマンドラインオプションを用いて、閾値を設定できます。例えば、閾値を10に設定するには、コマンドラインに`--sv-denovo-threshold 10`を追加します。閾値の初期設定値は20です。

DQスコアが閾値以上の場合、DNフィールドはDeNovoに設定されます。DQスコアが閾値未満の場合、DNフィールドはLowDQに設定されます。DQが0または「.」の場合、DQスコアは無効であり、DNフィールドは「.」に設定されます。

de novoスコアリングワークフロー

*de novo*構造多型スコアリングは、次のワークフローで使用できます。

- 2回のDRAGENランでの*de novo*スコアリングの実施。1回目では、pedigreeファイル内の全サンプルに対して生殖細胞系列構造多型解析を一緒に実施します。2回目では、*de novo*構造多型スコアリングをジョイント生殖細胞系列VCF出力に適用します。213 ページの 「2回のランのワークフロー」 を参照してください。
- 1回のDRAGENランでの*de novo*スコアリングの実施。pedigreeファイル内のすべてのサンプルに対して生殖細胞系列構造多型解析を一緒に実施し、その後、ジョイント生殖細胞系列構造多型コールに*de novo*スコアリングを適用します。214 ページの 「1回のランのワークフロー」 を参照してください。

2回のランのワークフロー

2回のランのワークフローでは、最初に、次の例に示すように、複数のサンプルに対して標準的なDRAGENのジョイント生殖細胞系列解析を実行します。

```
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_DIR1> \
--output-file-prefix <PREFIX1>
```

2回目のランでは、VCF出力（<OUT_DIR1>/<PREFIX1>.sv.vcf.gz）を*de novo*スコアリングのインプットとして使用します。--variantオプションを用いることで、VCFインプットを得ることができます。次のコマンドラインは、2回目のランの例を示します。

```
dragen -f \
--variant <MULTI_SAMPLE_VCF_FILE> \
--pedigree-file <PED_FILE> \
--enable-map-align false \
--sv-denovo-scoring true \
--output-directory <OUT_DIR2> \
--output-file-prefix <PREFIX2>
```

結果として生じる出力VCFファイル（<OUT_DIR2>/<PREFIX2>.sv.vcf.gz）には、*de novo*スコアリングアノテーションすべてが含まれます。

1回のランのワークフロー

必要な*de novo*スコアリングオプションすべてを用いて、複数のサンプルに対して標準的なDRAGENのジョイント生殖細胞系列解析を実施します。次の例は1回のランのワークフローを示します。

```
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 \
--pedigree-file <PED_FILE>
```

結果として生じる出力VCFファイル（`<OUT_DIR>/<PREFIX>.sv.vcf.gz`）には、*de novo*スコアリングアノテーションすべてが含まれます。

Ploidy Estimator

Ploidy Estimatorは初期設定で実行されます。Ploidy Estimatorは、マッパー/アライナーからのリードを用いて、ヒトゲノム中の各常染色体と異質染色体のカバレッジのシーケンス深度を計算します。次に、性染色体カバレッジ中央値と常染色体カバレッジ中央値との比率を用いて、サンプルの性核型を推定します。性核型はその比率の範囲に基づいて推定されます。この比率がすべての予想範囲外の場合、Ploidy Estimatorは性核型を決定しません。

性核型	最小X比率	最大X比率	最小Y比率	最大Y比率
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

ploidyの推定は、インプットシーケンスデータのタイプを決定できない場合や、常染色体にシーケンスカバレッジが十分にない場合に失敗することがあります。ploidyの推定が失敗した場合、推定カバレッジ中央値は0（ゼロ）となります。

腫瘍リードおよび一致した正常リードの両方がインプットとして提供される場合、Ploidy Estimatorは、一致した正常サンプルのシーケンスカバレッジおよび性核型のみを推定し、腫瘍リードを無視します。腫瘍リードのみがインプットとして提供された場合、Ploidy Estimatorは腫瘍サンプルのシーケンスカバレッジおよび性核型を推定します。

出力メトリクス

Ploidy Estimatorの結果は、ノーマライズしたコンティグごとのカバレッジ中央値を含めて、`<output-file-prefix>.ploidy_estimation_metrics.csv`ファイルおよび標準出力でレポートされます。結果の例を次に示します。

```
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
```

Ploidy Caller

Ploidy Callerは、Ploidy Estimatorから得たコンティグごとのカバレッジ中央値を用いて、全ゲノムシーケンスデータからヒト生殖細胞系列サンプル中の数的異常および染色体モザイク現象を検出します。

Ploidy Callerは、次の場合を除き、初期設定で実行されます：

- Ploidy Estimatorが、インプットデータが全ゲノムシーケンスからのものかどうかを判断することができない。例えば、エクソームまたはターゲットシーケンスからのデータなど。
- リファレンスゲノムに、ヒトに予想される22の常染色体と2つの異質染色体が含まれていない。
- 生殖細胞系列サンプルが存在しない。例えば、Tumor-only解析の場合。

コールモデル

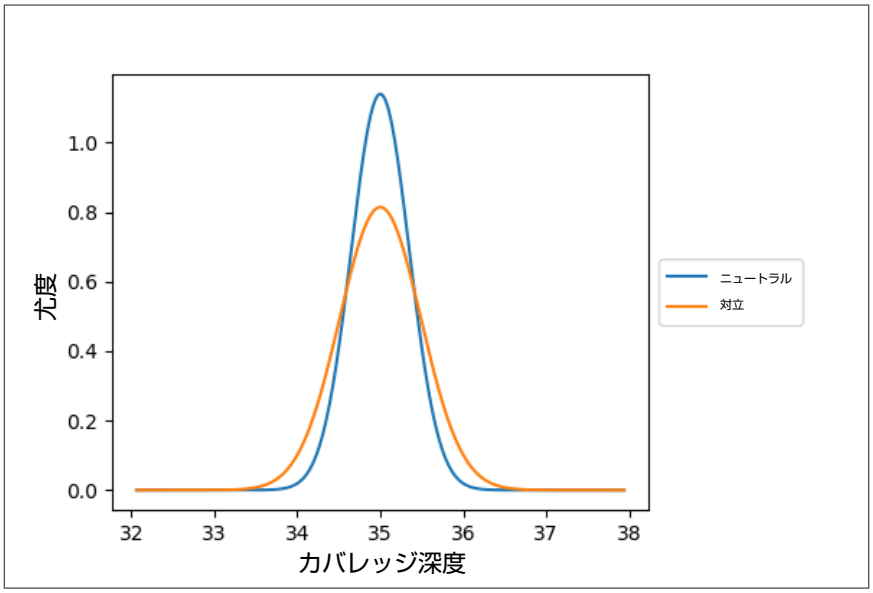
染色体モザイク現象は、常染色体全体のカバレッジ中央値と比較して、染色体のカバレッジ中央値に大きな変化がみられる場合に検出されます。

次の表は、所定の数的異常画分とモザイク画分について予想されるカバレッジの変化の例を示したものです。

ニュートラルコピー数	バリアントコピー数	モザイク画分	予想されるカバレッジ変化
2	1	10%	-5%
2	1	5%	-2.5%
2	3	5%	+2.5%
2	3	10%	+5%

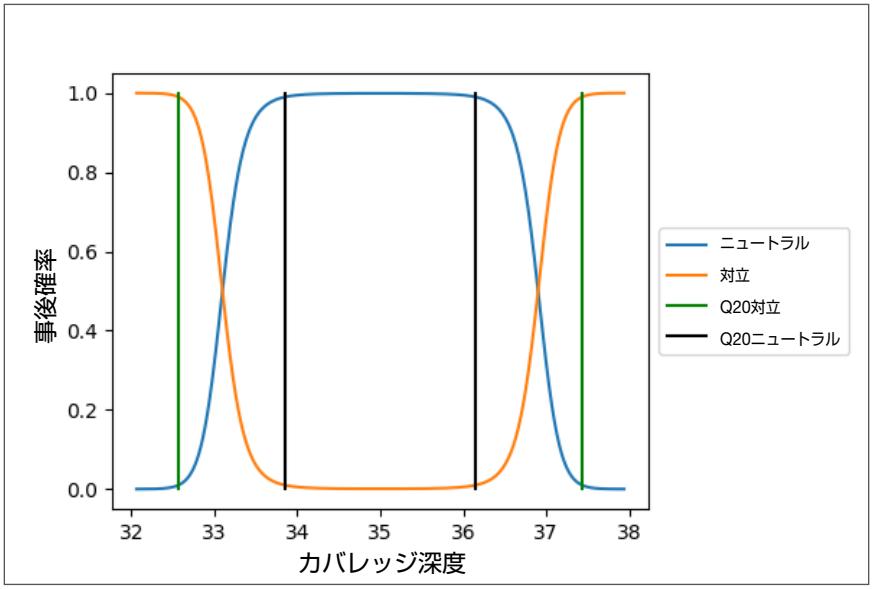
Ploidy Callerは、帰無（ニュートラル）仮説と対立（モザイク）仮説の両方について、カバレッジを正規分布としてモデル化します。2つの正規分布はサンプルの常染色体カバレッジ中央値で等しい平均を持ちますが、対立仮説の正規分布の分散は帰無仮説の正規分布の分散より大きくなります。30xカバレッジにおける2モデルのベースライン分散は、約2,500のWGSサンプルのコホートから経験的に決定されたものです。2モデルに使用される実際の分散は、30xカバレッジにおけるベースライン分散を、サンプルの常染色体カバレッジ中央値に対して調整して計算されます。常染色体の35xカバレッジ中央値のサンプルについて、帰無仮説と対立仮説の尤度分布を以下に示します。

図 11　帰無仮説と対立仮説の尤度分布



染色体モザイク現象に対して経験的に推定された事前確率を適用した後、Ploidy Callerは、常染色体の35xシーケンスカバレッジの中央値のサンプルについて以下に示すとおり、帰無仮説および対立仮説の事後確率に従ってploidyコールを生成します。

図 12　帰無仮説と対立仮説の事後確率



常染色体の35xカバレッジの中央値の場合、ニュートラル（REF）コールと対立（DELまたはDUP）コールのどちらかを決定する閾値は、常染色体のカバレッジのほぼ±5%のシフトです。常染色体の100xカバレッジの中央値の場合、その閾値は常染色体のカバレッジのほぼ+/0.3%のシフトです。Q20閾値は低品質コールのフィルタリングに使用されます。

Ploidy Callerリファレンス性核型

予想されるリファレンスploidyが2つの場合に、常染色体の数的異常と染色体モザイク現象を検出することに加えて、Ploidy Callerは異質染色体でもこれらのバリアントを検出することができます。

異質染色体のコールに使用されるリファレンス性核型は、`--sample-sex`を使用したコマンドライン、またはPloidy Estimatorのいずれかから提供されるサンプルの性核型から決定されます。サンプルの性核型がコマンドラインで提供されず、Ploidy Estimatorによって決定されない場合、性核型はXXと仮定されます。性核型が少なくとも1本のY染色体を含む場合、リファレンス性核型はXYです。性核型が少なくとも1本のY染色体を含まない場合、性核型はXXです。

次の表は、サンプルにおいて可能性のある性染色体をそれぞれ示したものです。Y染色体のリファレンスploidyが0（ゼロ）である場合、Y染色体ではploidyコールは行われません。

性核型	XリファレンスPloidy	YリファレンスPloidy
XX	2	0
XY	1	1
XXY	1	1
XYY	1	1

性核型	Xリファレンスploidy	Yリファレンスploidy
X0	2	0
XXXY	1	1
XXX	2	0

Ploidy Caller出力ファイル

Ploidy Callerは、出力ディレクトリに<output-file-prefix>.ploidy.vcf.gz出力ファイルを生成します。出力ファイルはVCF 4.2仕様に準拠しています。リファレンス常染色体および異質染色体については、それぞれ1つの記録がレポートされます。ただし、Y染色体については、リファレンス性核型がXXである場合を除きます。リファレンスゲノム中の他のシーケンス、例えばミトコンドリアDNA、局在していないシーケンスや配置されていないシーケンス、代替コンティグ、デコイコンティグ、エプスタインバーウイルスシーケンスなどについては、コールは行われません。

VCFファイルには、次の情報が含まれています。

- **メタ情報**：VCF出力ファイルには、`DRAGENVersion`や`DRAGEN CommandLine`などの一般的なメタ情報と、Ploidy Caller固有の情報が含まれます。VCFヘッダーには、常染色体カバレッジ深度中央値に関するメタ情報、提供された性核型（利用可能な場合）、Ploidy Estimatorから推定された性核型（利用可能な場合）、およびリファレンス性核型が含まれます。ヘッダー行の例を次に示します：

  ```
  ##autosomeDepthOfCoverage=36.635
  ##providedSexKaryotype=XY
  ##estimatedSexKaryotype=X0
  ##referenceSexKaryotype=XY
  ```

- **FILTERフィールド**：VCF出力ファイルには、クオリティスコアが20未満の結果をフィルタリングするLowQualフィルタが含まれています。
- **INFOフィールド**：VCF出力INFOフィールドには、次のものがあります：
  - END：この記録に記述されているバリアントの終了位置。
  - SVTYPE：構造多型のタイプ。
- **Formatフィールド**：VCF出力ファイルには、次のフォーマットフィールドが含まれます。`GT` FORMATフィールドはありません。VCFのバリアントコールでは、ALT列に`<DUP>`または`<DEL>`が表示されます。ALT列には、非バリアントコールが表示されます。出力ファイルを下流で使用する場合、GTフィールドを追加することができます。バリアントコールの場合、二倍体コンティグには./1を用い、一倍体コンティグには`1`を用います。非バリアントコールの場合、二倍体には`0/0`、一倍体には`0`を使用します。
  - DC：カバレッジ深度。
  - NDC：ノーマライズしたカバレッジ深度。

以下は出力ファイルの例です。

```
##fileformat=VCFv4.2
```

```
...
##autosomeDepthOfCoverage=36.635
##providedSexKaryotype=XY
##estimatedSexKaryotype=X0
##referenceSexKaryotype=XY
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the
variant described in this record">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural
variant">
##ALT=<ID=DEL,Description="Deletion relative to the reference">
##ALT=<ID=DUP,Description="Region of elevated copy number relative to the
reference">
##FILTER=<ID=LowQual,Description="QUAL below 20">
##FORMAT=<ID=DC,Number=1,Type=Float,Description="Depth of coverage">
##FORMAT=<ID=NDC,Number=1,Type=Float,Description="Normalized depth of
coverage">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT MySampleName
chr1 1 . N . 31.1252 PASS END=248956422 DC:NDC 36.836:1.00549
chr2 1 . N . 31.451 PASS END=242193529 DC:NDC 36.668:1.0009
...
chr21 1 . N . 31.4499 PASS END=46709983 DC:NDC 36.6:0.999045
chr22 1 . N . 28.8148 PASS END=50818468 DC:NDC 37.2:1.01542
chrX 1 . N . 29.7892 PASS END=156040895 DC:NDC 18:0.982667
chrY 1 . N <DEL> 150 PASS END=57227415;SVTYPE=DEL DC:NDC 5.7:0.311178
```

細胞株アーティファクト

細胞株由来のサンプルには、一部の染色体上でバリアントploidyコールとなる可能性のあるカバレッジアーティファクトが含まれることが多くあります。染色体17、19および22は、細胞株カバレッジアーティファクトで最も一般的です。細胞株サンプルについてploidyコールの精度評価を行う場合は、既知の細胞株アーティファクトを有する染色体をフィルタリングで除外します。

QCメトリクスおよびカバレッジ/コール可能性レポート

DRAGENは、各実行中にパイプライン固有のメトリクスカバレッジレポートを生成します。パイプラインの異なるステージで生成されるメトリクスには、次の4つのグループがあります：

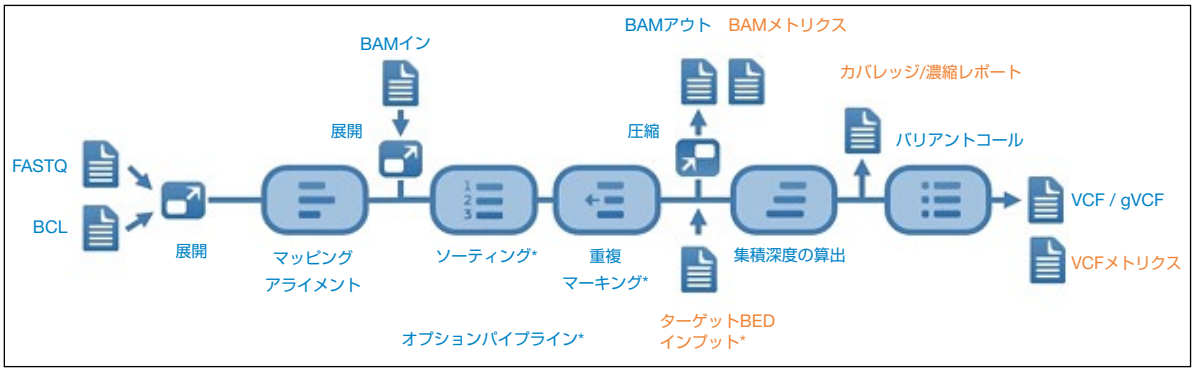
- マッピングおよびアライメントのメトリクス
- VCFメトリクス
- 期間（またはランタイム）メトリクス

- カバレッジ（または濃縮）メトリクスとレポート

マッピング/アライメントメトリクス、VCFメトリクス、期間メトリクス、および利用可能なカバレッジレポートのサブセットは自動生成されるため、アクティブ化や特定のコマンドは必要ありません。追加のカバレッジメトリクスを有効にしたり、追加のカバレッジ領域を指定したりできます。

DRAGENは解析中にメトリクス計算を実行し、ランタイムに影響を与えないようにします。

図 13　メトリクスとレポートの生成



QCメトリクス出力フォーマット

QCメトリクスは標準出力に出力され、CSVファイルは実行出力ディレクトリに書き込まれます。

- <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

セクション	RG/サンプル	メトリクス	カウント/比率/時間	パーセント/秒
MAPPING/ALIGNING SUMMARY		全インプットリード	816360354	
MAPPING/ALIGNING SUMMARY		重複リード数（削除されていないとマークされたもの）	15779031	1.93

...

セクション	RG/サンプル	メトリクス	カウント/比率/時間	パーセント/秒
MAPPING/ALIGNING PER RG	RGID_1	RG内のリード数合計	816360354	100
MAPPING/ALIGNING PER RG	RGID_1	重複リード数（マークされたもの）	15779031	1.93
...				
VARIANT CALLER SUMMARY		サンプル数	1	
VARIANT CALLER SUMMARY		処理されたリード数	738031938	
...				
VARIANT CALLER PREFILTER	サンプル_1	合計	4918287	100
VARIANT CALLER PREFILTER	サンプル_1	二対立遺伝子	4856654	98.75
...				
RUN TIME		リファレンスのローディング時間	00:18.6	18.65
RUN TIME		リードのアライメント時間	19:24.4	1164.42

マッピングおよびアライメントのメトリクス

SAMtools Flagstatコマンドによって計算されたメトリクスなどのマッピングおよびアライメントのメトリクスは、集約レベル（インプットデータ全体）およびリードグループごとのレベルで入手可能です。特に明記されていない限り、メトリクス単位はリードです（すなわち、ペアやアライメント単位ではなく）。

メトリクス	説明
Total input reads	インプットFASTQファイルのリードの合計数。
Number of duplicate marked reads	`--enable-duplicate-marking`オプションをtrueに設定した結果、重複としてマークされたリード。

メトリクス	説明
Number of duplicate marked and mate reads removed	重複としてマークされたリードと、`--remove-duplicates`オプションがtrueに設定されている場合に削除されるメイトリード。
Number of unique reads	リードの合計数から重複してマークされたリードを引いた数。
Reads with mate sequenced	メイトのあるリード数。
Reads without mate sequenced	リードの合計数からメイトシーケンスされたリード数を引いたもの。
QC-failed reads	プラットフォーム/ベンダーのクオリティチェックに合格しなかったリード（SAMフラグ0x200）。
Mapped reads	マッピングされたリードの合計数からマッピングされていないリードの数を引いた数。
Number of unique and mapped reads	マッピングされたリードの数から重複マークされたリードの数を引いた数。
Unmapped reads	マッピングできなかったリードの合計数。
Singleton reads	リードをマッピングできたが、ペアになるメイトを読み取ることができなかったリード数。
Paired reads	ペアの両方のリードがマッピングされているリードのカウント。
Properly paired reads	ペア内の両方のリードはマッピングされ、推定された挿入長分布に基づいて互いに許容範囲内に収まります。
Not properly paired reads (discordant)	ペアリードの数から、適切にペアになったリードの数を引いた数。
Paired reads mapped to different chromosomes	メイトのあるリードの数（メイトは別の染色体にマッピングされている場合）。
Paired reads mapped to different chromosomes (MAPQ >= 10)	MAPQ>10で、メイトのあるリード数（メイトは別の染色体にマッピングされている場合）。
Reads with indel R1	少なくとも1つのindelを含むR1リードのパーセンテージ。
Reads with indel R2	少なくとも1つのindelを含むR2リードのパーセンテージ。
Soft-clipped bases R1	ソフトクリッピングされたR1リード中の塩基のパーセンテージ。
Soft-clipped bases R2	ソフトクリッピングされたR2リード中の塩基のパーセンテージ。

メトリクス	説明
Mismatched bases R1	R1上のミスマッチした塩基数。これはSNPカウントとindel長の合計です。このメトリクスはソフトクリッピングやRNAイントロン内のものは何もカウントしません。また、リファレンス塩基またはリード塩基のいずれかがNの場合、メトリクスは、ミスマッチをカウントしません。
Mismatched bases R2	R2上のミスマッチした塩基数。これはSNPカウントとindel長の合計です。このメトリクスはソフトクリッピングやRNAイントロン内のものは何もカウントしません。また、リファレンス塩基またはリード塩基のいずれかがNの場合、メトリクスは、ミスマッチをカウントしません。
Mismatched bases R1 (excluding indels)	R1上のミスマッチ塩基数。indelの長さは無視されます。ソフトクリッピングやRNAイントロン内のものは何もカウントされません。また、リファレンス塩基またはリード塩基のいずれかがNの場合、メトリクスは、ミスマッチをカウントしません。
Mismatched bases R2 (excluding indels)	R2上のミスマッチ塩基数。indelの長さは無視されます。このメトリクスはソフトクリッピングやRNAイントロン内のものは何もカウントしません。また、リファレンス塩基またはリード塩基のいずれかがNの場合、メトリクスは、ミスマッチをカウントしません。
Q30 Bases	BQが30以上の塩基数合計。
Q30 Bases R1	R1上のBQが30以上の塩基数合計。
Q30 Bases R2	R2上のBQが30以上の塩基数合計。
Q30 Bases (excluding dups and clipped bases)	BQが30以上の非重複塩基およびクリップされていない塩基数。
Histogram of reads map qualities	• MAPQ [40:inf]のリード • MAPQ [30:40)のリード • MAPQ [20:30)のリード • MAPQ [10:20)のリード • MAPQ [0:10)のリード
Total alignments	0を超えるクオリティでアライメントされた座位のリード数合計。
Secondary alignments	二次アライメント座位の数。
Supplementary (chimeric) alignments	キメラリードは複数の座位にまたがって分裂します（おそらく構造多型による）。1つのアライメントは代表アライメントと呼ばれます。もう1つは補足的なものです。
Estimated read length	インプット塩基の合計数をリード数で除したもの。
Histogram	233 ページの「ヒストグラムカバレッジレポート」を参照してください。

メトリクス	説明
PCT of bases aligned that fell inside the interval region	間隔領域とターゲット領域内の塩基数をアライメントした塩基数合計で除した値。
Estimated sample contamination	別のヒト由来である可能性のあるサンプル中のリードの推定割合。

バリアントコーリングの予測精度はクロスサンプルコンタミネーションの影響を受けます。特に、アリル頻度の低いバリアントを検出することを目的とするパイプラインでは、少量のコンタミネーションであっても多くのFP コールにつながる可能性があります。

DRAGEN クロスサンプルコンタミネーションモジュールは、確率的混合モデルを用いて、別のヒト由来である可能性のあるサンプル中のリードの割合を推定します。このサンプルコンタミネーション割合は、複数のパイルアップ位置で観察されるリードの尤度を最大化する混合モデルにおけるパラメーター値として推定されます。この混合モデルは集団アリル頻度と推定されるサンプル遺伝型を説明します。

生殖細胞系列モードでこのメトリクスを有効にするには、集団アリル頻度を持つマーカー座位（RSID）を含む VCF へのファイルパスをコマンドラインで指定する必要があります。

```
--qc-cross-cont-vcf /opt/edico/config/sample_cross_
contamination_ resource_[hg19 or GRCh37 or GRCh38].vcf
```

体細胞モードでは、コンタミネーションアルゴリズムはまず、CNV または LOH によって導入され得るバイアスを回避しようとします。このアルゴリズムは、FFPE サンプルについて調整するために、サンプルからのヌクレオチドノイズも推定します。

体細胞コンタミネーション検出を有効にするには、次の設定を使用します。

```
--qc-somatic-contam-vcf /opt/edico/config/somatic_sample_cross_
contamination_resource_[hg19 or GRCh37 or GRCh38].vcf.gz
```

DRAGEN に付属する VCF リソースファイルは、Ensembl データベースから再構築することができます。DRAGEN のコンティグフォルダーに含まれる VCF ファイルには、集団 AF が 0.5 に近いマーカー位置約 5,000 が含まれます。ファイルはリファレンス固有です（hg19/GRCh37/hg38）。互換性のないリソースとリファレンスファイル（例：CRCh37 リソースファイルと hg19 リファレンス）が使用された場合、DRAGEN は中止されます。

次に、推定コンタミネーションが 1.1% のサンプルの出力例を示します。この値は少数で提供されるため、値 0.011 は 1.1% と同じです。

```
MAPPING/ALIGNING SUMMARY Estimated sample contamination 0.011
```

体細胞モードマッピングおよびアライメントのメトリクス

体細胞Tumor-Normalモードでは、マッピングおよびアライメントのメトリクスは腫瘍サンプルおよび正常サンプルに対して別々に生成され、各行はTUMORまたはNORMALで始まり、当該サンプルを示します。腫瘍サンプルのメトリクスが最初に出力され、次に正常サンプルのメトリクスが出力されます。

リードグループごとのメトリクスも、腫瘍リードグループと正常リードグループに分けられます。

体細胞Tumor-onlyのモードでは、マッピングおよびアライメントのメトリクスは生殖細胞系列モードと同じ規則に従い、TUMORまたはNORMALとラベル付けされたメトリクスはありません。

バリアントコールのメトリクス

生成されるバリアントコールメトリクスは、RTG vcfstatsによって計算されたメトリクスと類似しています。メトリクスは、複数サンプルのVCFファイルおよびgVCFファイルで各サンプルについてレポートされます。実行ケースに基づいて、メトリクスは標準のVARIANT CALLERまたはJOINT CALLERとしてレポートされます。メトリクスは、未処理のVCFファイル（PREFILTER）とハードフィルター処理されたVCFファイル（POSTFILTER）の両方についてレポートされます。

正常サンプルのパネル（PON）およびCOSMICフィルターされたバリアントは、POSTFILTER VCFメトリクスではPASSバリアントとしてカウントされます。これらのPASSバリアントが原因で、POSTFILTER VCFメトリクスではバリアントカウントが予想よりも多くなる可能性があります。

メトリクス	説明
Number of samples	集団/ジョイントVCF内のサンプル数
Reads Processed	バリアントコールに使用されるリード数。重複してマークされたリードと、ターゲット領域外にあるリードを除く。
Total	バリアントの合計数（SNP+MNP+indel）。
Biallelic	観察されたアリルを2つ含むゲノム中の部位数。リファレンスは1アリルとしてカウントされ、バリアントアリルが1つ許容されます。
Multiallelic	観察されたアリルを3つ以上含むVCF中の部位数。リファレンスは1としてカウントされ、バリアントアリルが2つ以上許容されます。
SNPs	リファレンス、アリル1およびアリル2がすべて長さ1である場合、バリアントはSNPとしてカウントされます。
Insertions (Hom)	ホモ接合性挿入を含むバリアント数。
Insertions (Het)	両方のアリルが挿入であるがホモ接合性ではないバリアント数。
Deletions (Het)	ホモ接合性欠失を含むバリアント数。
INDELS (Het)	遺伝型が以下のいずれかであるバリアント数 [挿入+欠失]、[挿入+SNP]、または[欠失+SNP]

メトリクス	説明
De Novo SNPs	DQが0.05を上回り、*de novo*とマークされたSNP。*--qc-snp-denovo-quality-threshold*オプションを必要な閾値に設定します。初期設定値は0.05です。
De Novo INDELs	DQ値が0.02を上回り、*de novo*マーキングされたindel。このDQ閾値は、*--qc-indel-denovo-quality-threshold*オプションを必要なDQ閾値に設定することで指定できます。初期設定値は0.02です。
De Novo MNPs	DQが0.05を上回り、*de novo*とマークされたSNP。*--qc-snp-denovo-quality-threshold*を必要な閾値に設定します。初期設定値は0.05です。
(Chr X SNPs)/(Chr Y SNPs) ratio in the genome (or the target region)	染色体X（または染色体Xとターゲット領域との交点）のSNP数を、染色体Y（または染色体Yとターゲット領域との交点）のSNP数で除した値。X染色体またはY染色体のいずれにもアライメントがなければ、このメトリクスはNAと表示されます。
SNP Transitions	2つのプリン（A<->G）または2つのピリミジン（C<->T）の交換。
SNP Transversions	プリン塩基とピリミジン塩基の交換Ti/Tv比:遷移と遷移の比。
Heterozygous	ヘテロ接合性バリアント数。
Homozygous	ホモ接合性バリアント数。
Het/Hom ratio	ヘテロ接合性/ホモ接合性の比。
In dbSNP	dbSNPリファレンスファイル内に存在する検出されたバリアント数。*--bsnp*オプションを使用してdbSNPファイルを指定しない場合、`In dbSNP`と`Novel`の両方のメトリクスはNAと表示されます。
Novel	バリアント数の合計からdbSNP中のバリアント数を引いたもの。
Percent Callability	生殖細胞系列モードおよび体細胞系モードで利用でき、gVCF出力を伴います。PASSするジェノタイプコールをもつ非Nリファレンス位置の割合。マルチアリルバリアントはカウントされません。欠失は、ホモ接合性コールの場合のみ、すべての欠失リファレンス位置についてカウントされます。常染色体と染色体X、Y、およびMのみを考慮します。
Percent Autosome Callability	常染色体のみを考慮します。
Percent QC Region Callability in Region i (i is equivalent to regions 1, 2, or 3)	*--qc-coverage-region-i*オプションを使用してカスタム領域のコール可能性を要求し、コール可能性の出力を*--qc-coverage-reports-i*で指定する場合に使用できます。すべてのコンティグを考慮します。

コンティグあたりのHet/Hom比

生殖細胞系列スモールバリアントコーラーが実行されると、DRAGENはコンティグあたりのヘテロ接合性/ホモ接合性比を計算します。DRAGENは、未処理のVCF（PREFILTER）とハードフィルター処理された（POSTFILTER）VCFの両方の比率をレポートします。メトリクスは、`.vc_hethom_metrics.csv`ファイルに出力されます。このファイルには、処理された各一次コンティグの次の値が含まれています。

- コンティグ
- ヘテロ接合性バリアント数
- ホモ接合性バリアント数
- ヘテロ接合性/ホモ接合性（Het/Hom）比

次の例は、メトリクスのセクションを示します。

```
VARIANT CALLER POSTFILTER,HG04070,1 Heterozygous,185733
VARIANT CALLER POSTFILTER,HG04070,1 Homozygous,182928
VARIANT CALLER POSTFILTER,HG04070,1 Het/Hom ratio,1.015
VARIANT CALLER POSTFILTER,HG04070,2 Heterozygous,203946
VARIANT CALLER POSTFILTER,HG04070,2 Homozygous,174294
VARIANT CALLER POSTFILTER,HG04070,2 Het/Hom ratio,1.170
VARIANT CALLER POSTFILTER,HG04070,3 Heterozygous,192861
VARIANT CALLER POSTFILTER,HG04070,3 Homozygous,130087
VARIANT CALLER POSTFILTER,HG04070,3 Het/Hom ratio,1.483
VARIANT CALLER POSTFILTER,HG04070,4 Heterozygous,178389
VARIANT CALLER POSTFILTER,HG04070,4 Homozygous,157062
VARIANT CALLER POSTFILTER,HG04070,4 Het/Hom ratio,1.136
```

Het/Hom比の値は、染色体全体の片親性ダイソミー（UPD）の指標として用いることができます。ある種の染色体のUPDはインプリンティング異常症として知られる遺伝的症候群と関連しています。全染色体UPDのHet/Hom比は0.0に近い値です。範囲はさまざまですが、通常は1.0～2.0です。Het/Hom比を自分のシーケンスデータと照らし合わせて解釈していることを確認してください。

期間メトリクス

期間メトリクスセクションには、各プロセスのランタイムの内訳が含まれます。例えば、次のメトリクスは、マッパーとバリアントコーラーのパイプラインに対して生成されます：

- リファレンスのローディング時間
- リードのアライメント時間
- ソーティングと重複マーキングの時間
- DRAGStrキャリブレーション時間
- 部分再構成時間

- バリアントコール時間
- 合計ランタイム

コール可能性レポート

DRAGENは、gVCF出力モードでスモールバリアントコーラーを実行すると、バリアントコーラーメトリクスの一部としてコール可能性レポートを自動的に生成します。DRAGENは、テストで指定されたBEDファイルに含まれる各領域のコール可能性の割合（%）を追加し、BEDフォーマットとCSVフォーマットの両方で完全なコール可能性レポートを生成します。コール可能性は、PASSするジェノタイプコールを有する非Nリファレンス位置の割合として定義されます。コール可能性メトリクスの計算には、次の基準が使用されます：

- gVCF全体のコール可能性が計算されます。
- マルチアリルバリアントはカウントされません。
- デコイコンティグは無視されます。
- 未配置および未ローカライズのコンティグは無視されます。
- Nの位置は、コール不可と見なされます。
- バリアントコールが実行されなかった領域では、コール可能性は0です。
- ホモ接合性欠失は、その欠失によってスパンするリファレンス位置すべてに対する、PASSするジェノタイプコールとしてカウントされます。

*--vc-target-bed*オプションを指定すると、出力は`target_bed_callability.bed`ファイルになり、このファイルには、インプットターゲットBED領域全体のコール可能性と常染色体コール可能性が含まれます。*--vc-target-bed-padding*オプションで指定されたパディングサイズが使用され、重複する領域がマージされます。

コール可能性については、カスタム領域カバレッジ/コール可能性レポートとともに出力することもできます。

カスタム領域全体のカバレッジ/コール可能性レポート

DRAGENでは、次のカバレッジレポートが生成されます：

- *--vc-target-bed*オプションが指定されている場合は、ゲノム全体またはターゲット領域の初期設定レポート一式。
- 必要に応じた、最大3つの対象領域（カバレッジ領域）の追加レポート。
  DRAGEN は、指定された領域ごとに、初期設定のレポートと、その領域に要求されたオプションのレポートを生成します。

カバレッジ領域レポートを生成するには、*–qc-coverage-region-i*オプション（iは1、2、または3）を使用します。

- 各*–qc-coverage-region-i*オプションには、BEDファイル引数が必要です。
- 各BEDファイル内の領域は、オプションで*--qc-coverage-region-padding-i*オプション（初期設定値は0）を使用してパディングすることができます。
- 初期設定レポート一式は、領域ごとに生成されます。
- オプションで、*–qc-coverage-reports-i* iオプションを用いると、領域ごとに追加のレポートを指定できます。

次の例は、カバレッジレポートの生成に必要なオプションを示しています。

```
$ 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
```

リードと塩基のカウント

表8および表 9にリストされているすべての初期設定レポートおよびオプションのカバレッジレポートでは、リードおよび塩基のカウントに次の初期設定ルールが使用されます。

- 重複リードは無視されます。
- ソフトクリップおよび/またはハードクリップされた塩基は無視されます。
- 重複するメイトは二重にカウントされます。
- MAPQ > 0のリードが含まれます。MAPQ = 0のリードがフィルタリングされます。
- BQ ≥ 0は含まれます。

初期設定されない設定：

レポートは、マッパーやアライナー、またはバリアントコーラーを実行してもしなくても利用できます。ただし、*--enable-sort*オプションはtrue（初期設定はtrue）に設定する必要があります。

初期設定では、重複するメイトは二重にカウントされます。*--qc-coverage-ignore-overlaps=true*と設定して、各断片のアライメントをすべて解決し、重複する塩基を二重にカウントしないようにします。これにより、ランタイムが若干長くなる可能性があります。このオプションでは、*--enable-map-align=true*も設定する必要があります。*--gc-coverage-ignore-overlaps*はグローバル設定であり、すべてのqc-coverage-reportsを更新します。

初期設定では、ソフトクリップされた塩基はカバレッジにカウントされません。*--qc-coverage-count-soft-clipped-bases=true*を設定すると、ソフトクリップされたベースがカバレッジ計算に含まれます。*--qc-coverage-count-soft-clipped-bases*はグローバル設定であり、すべてのqc-coverageレポートを更新します。

領域ごとにオプションのレポートを任意に組み合わせることができます。領域ごとに複数のレポートタイプを選択する場合は、スペースで区切る必要があります。

qc-coverage-filtersを使用して、最小MAPQおよび最小BQを上書きして、特定の領域に適用することができます。

カバレッジフィルターを有効にするには、*--qc-coverage-filters-i*オプション（iは1、2、または3）のいずれかを、関連する*--qc-coverage-region-i*オプションと組み合わせて使用します。`--qc-coverage-filters-i`の初期設定値は`mapq<1`、`bq<0`です。初期設定ではすべてのBQが含まれますが、MAPQ = 0のリードはフィルタリングされます。

- *--qc-coverage-region-i=<targetedregions.bed>*
- *--qc-coverage-filters-i <filters string>*

例えば、次のオプションを使用して、1 bpの解像度のカバレッジ出力とフィルタリングを有効にします：

```
--qc-coverage-region-1 <targetedregions.bed>
```

```
--qc-coverage-filters-1 'mapq<10,bq<30'
--qc-coverage-reports-1 full_res
```

- 引数の構文はmapq<value,bq<valueです。これは、マッピングクオリティが指定値より低いリードはカウントされないこと、および/またはベースコールクオリティが指定値より低い塩基はカウントされないことを意味します。
- 有効なフィルター引数はmapqおよびbqのみです。いずれか、または両方を指定できます。
- 1つの演算子<のみがサポートされています。<=、>、>=、=はサポートされていません。
- フィルタリングが対象の領域に対して有効になっている場合、DRAGENはこの領域のフィルタリングされたレポートファイルを出力します。フィルタリングされていないレポートファイルは、フィルタリングされたターゲット領域に出力されません。

カスタム領域のコール可能性

*--qc-coverage-region-i*オプションを*--qc-coverage-reports-i*（iは1、2、または3）とともに使用する場合、コール可能性をその領域のレポートタイプとして追加できます。出力は`qc-coverage-region-i-callability.bed`ファイルです。指定した`qc-coverage-region-i`ファイルごとに、コール可能性平均がバリアントコールメトリクスファイルでレポートされます。*--qc-coverage-region-padding-i*で指定されたパディングサイズが使用され、重複する領域がマージされます。

オプションの最小MAPQフィルターと最小BQフィルターは、リードカウントと塩基カウントだけに影響し、コール可能性レポートには影響しません。

コンティグの長さおよび関心領域の長さ（分母として使用）には、FASTAにNをもつ領域は含まれません。

利用可能なレポートタイプ

表8　初期設定レポート

ファイル名	説明
_coverage_metrics.csv	231 ページの 「カバレッジメトリクスレポート」
_fine_hist.csv	233 ページの「詳細ヒストグラムカバレッジレポート」
_hist.csv	233 ページの「ヒストグラムカバレッジレポート」
_overall_mean_cov.csv	234 ページの「全体平均カバレッジレポート」
_contig_mean_cov.csv	234 ページの「コンティグあたりの平均カバレッジレポート」
_ploidy.csv	304 ページの「出力メトリクス」

ファイル名	説明
_read_cov_report.bed	238 ページの「リードカバレッジレポート」

表 9　オプションのレポート

ファイル名	説明
_full_res.bed	234 ページの「Full Resレポート」
_cov_report.bed	235 ページの「カバレッジレポート」
_callability.bed	228 ページの「コール可能性レポート」

カバレッジメトリクスレポート

カバレッジメトリクスレポートでは、`_coverage_metrics.csv`ファイルが出力されます。このファイルは、領域全体のメトリクスを提供します。領域には、ゲノム、ターゲット領域、QCカバレッジ領域があります。出力ファイルの最初の列にはCOVERAGE SUMMARY（カバレッジの要約）というセクション名が含まれ、2番目の列はすべてのメトリクスについて空です。

カバレッジを計算するときは、次の基準が使用されます：

- 重複したリードとクリッピングされた塩基は無視されます。
- 初期設定では、MAPQ <1のリード値とBQ <0の塩基は無視されます。`qc-coverage-filters-n`オプションを使用して、フィルタリングで除外するBQ塩基およびMAPQリードを指定できます。

次の表に、計算されたメトリクスを示します：

ファイル名	説明
Aligned bases in region	領域にユニークにマッピングされた塩基数と、ゲノムにユニークにマッピングされた塩基数に対する割合。
Average alignment coverage over region	領域にユニークにマッピングされた塩基数を領域中の部位数で割ったもの。
Uniformity of coverage (PCT > 0.2*mean) over region	領域の平均カバレッジの20%を超えるカバレッジを有する部位の割合。
PCT of region with coverage [ix, inf)	少なくともixのカバレッジの領域内の部位割合。iは100、50、20、15、10、3、1、または0に等しくなりえます。
PCT of region with coverage [ix, jx)	少なくともixで、jx未満のカバレッジの領域内の部位の割合。(i, j)は(50,100)、(20,50)、(15,20)、(10,15)、(3,10)、(1,3)、または(0,1)に等しくなりえます。

ファイル名	説明
Average chromosome X coverage over region	X染色体と領域の交点にアライメントした塩基数合計を、X染色体と領域の交点にある座位の合計数で除したもの。リファレンスゲノムに染色体Xが存在しない場合、または領域が染色体Xと交差していない場合、このメトリクスはNAと表されます。
Average chromosome Y coverage over region	Y染色体と領域の交点にアライメントした塩基数合計を、Y染色体と領域の交点にある座位の合計数で除したもの。リファレンスゲノムに染色体Yが存在しない場合、または領域が染色体Yと交差していない場合、このメトリクスはNAと表されます。
XAvgCov/YAvgCov ratio over genome/ target region	領域内のX染色体アライメントカバレッジ平均を領域内のY染色体アライメントカバレッジ平均で除したもの。リファレンスゲノムに染色体Xまたは染色体Yが存在しない場合、または領域が染色体Xまたは染色体Yと交差していない場合、このメトリクスはNAと表されます。
Average mitochondrial coverage over region	ミトコンドリア染色体と領域の交点にアライメントした塩基数合計を、ミトコンドリア染色体と領域の交点にある座位の合計数で除したもの。リファレンスゲノムにミトコンドリア染色体が存在しない場合、または領域がミトコンドリア染色体と交差していない場合、このメトリクスはNAと表されます。
Average autosomal coverage over region	領域内の常染色体座位にアライメントした塩基数合計を、領域内の常染色体座位の合計数で除したもの。リファレンスゲノム中に常染色体が存在しない場合、あるいはその領域が常染色体と交差していない場合、このメトリクスはNAと表されます。
Median autosomal coverage over region	領域内の常染色体座位全体のアライメントカバレッジ中央値。リファレンスゲノム中に常染色体が存在しない場合、あるいはその領域が常染色体と交差していない場合、このメトリクスはNAと表されます。
Mean/Median autosomal coverage ratio over region	領域内の常染色体の平均カバレッジを領域内の常染色体カバレッジ中央値で除したもの。リファレンスゲノム中に常染色体が存在しない場合、あるいはその領域が常染色体と交差していない場合、このメトリクスはNAと表されます。
Aligned reads in region	領域に一意にマッピングされたリードの数、およびゲノムにユニークにマッピングされたリードの数に対する割合（%）。MAPQが1以上のリードのみが含まれます。2次アライメントおよび補足アライメントは無視されます。

_coverage_metrics.csvファイルの内容の例を次に示します：

```
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
```

詳細ヒストグラムカバレッジレポート

詳細ヒストグラムレポートでは、_fine_hist.csvファイルが出力されます。このファイルには次の2つの列が含まれます：深度および全体。「深度」列の値は0から1000+の範囲であり、「全体」列は対応する深度でカバーされる座位の数を示します。

ヒストグラムカバレッジレポート

ヒストグラムレポートは_hist.csvファイルを出力します。このファイルには、次の情報が含まれます：

- カバレッジBED/ターゲットBED/WGS領域内で、特定のカバレッジ範囲内にある塩基の割合（%）。
- DRAGENを*--enable-duplicate-marking true*で実行した場合、重複リードは無視されます。

次の範囲が使用されます：

```
"[100x:inf)", "[1x:3x)", "[0x:1x)"
```

全体平均カバレッジレポート

全体平均カバレッジレポートでは、_overall_mean_cov.csvファイルが生成されます。このファイルには、カバレッジBED/ターゲットBED/WGS全体の平均アライメントカバレッジが含まれます（必要に応じて）。

_overall_mean_cov.csvファイルの内容の例を次に示します：

```
Average alignment coverage over target_bed,80.69
```

コンティグあたりの平均カバレッジレポート

コンティグ平均カバレッジレポートでは、_contig_mean_cov.csvファイルが生成されます。このファイルには、全コンティグの推定カバレッジと常染色体の推定カバレッジが含まれます。このファイルには、次の3つの列があります：

列1	列2	列3
コンティグ名	そのコンティグにアライメントした塩基の数。重複してマークされたリード、MAPQ=0のリードからの塩基、およびクリップされた塩基は除外されます。	次のように計算された推定カバレッジ。コンティグにアライメントした塩基の数（すなわち、列2）をコンティグの長さ、または（ターゲットBEDが用いられる場合は）そのコンティグをスパンするターゲット領域の全長で除したもの。

Full Resレポート

Full Resレポートは、_full_res.bedファイルをタブ区切りフォーマットで出力します。最初の3列は標準のBEDフィールドで、4列目は深度です。ファイル内の各レコードは、一定の深度を持つ指定された間隔のレコードです。深度が変更されると、新しいレコードがファイルに書き込まれます。マッピングクオリティ値が0、重複リード、およびクリッピングされた塩基を持つアライメントは、深度に対してカウントされません。

_full_res.bed出力ファイルには、ユーザー指定のカバレッジ領域のBED領域に該当する塩基の位置のみが表示されます。

_full_res.bedファイルの構造は、bedtools genomecov-bgの出力ファイルに類似しています。マッピング精度値が0のアライメントをフィルタリングで除外した後、およびターゲットBED（指定されている場合）によってフィルタリングした後に、bedtoolsコマンドラインを実行した場合、内容は同一です。

_full_res.bedファイルの内容の例を次に示します。

```
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
```

カバレッジレポート

`cov_report`レポートでは、BEDフォーマットとCSVフォーマットの両方で、タブ区切りのフォーマットされたカバレッジレポートファイルに`_cov_report.bed`ファイルが生成されます。最初の3列は標準のBEDフィールドです。残りの列フィールドは、同じレコード行で指定された間隔領域で計算された統計情報です。

次の表に、追加された列を示します。

列	説明
total_cvg	カバレッジ値の合計。
mean_cvg	平均カバレッジ値。
Q1_cv	下位四分位(25パーセンタイル)カバレッジ値。
median_cvg	カバレッジの中央値。
Q3_cvg	上位四分位(75パーセンタイル)のカバレッジ値。
min_cvg	最小カバレッジ値。
max_cvg	最大カバレッジ値。
pct_above_X	指定された間隔領域における、Xより大きい深度カバレッジを持つ塩基の割合(%)を示します。

初期設定では、間隔のカバレッジ合計が0の場合、レコードは出力ファイルに書き込まれます。カバレッジが0の間隔をフィルタリングで除外するには、構成ファイルで*vc-emit-zero-coverage-intervals*をfalseに設定します。

_cov_report.bedファイルの内容例を次に示します：

```
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
…
```

カバレッジ/コール可能性レポートの使用事例と期待される出力

次の表は、初期設定オプション（*--vc-target-bed*）とオプションのカバレッジ領域オプション（*--coverage-region*）を使用した場合に生成される出力を示しています。

--vc-target-bedを指定しますか? はい/いいえ	**--qc-coverage-region-iを指定しますか（iは1、2、3のいずれか）? はい/いいえ**	**期待される出力ファイル**
いいえ	いいえ	wgs_coverage_metrics.csv wgs_fine_hist.csv wgs_hist.csv wgs_overall_mean_cov.csv wgs_contig_mean_cov.csv

--vc-target-bedを指定しますか? はい/いいえ	--qc-coverage-region-iを指定しますか (iは1、2、3のいずれか)? はい/いいえ	期待される出力ファイル
いいえ	はい	wgs_coverage_metrics.csv wgs_fine_hist.csv wgs_hist.csv wgs_overall_mean_cov.csv wgs_contig_mean_cov.csv ユーザーが指定したカバレッジ領域ごとに: 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(full_resレポートタイプがqc-coverage-region-iに要求された場合) qc-coverage-region-i_cov_report.bed(cov_reportレポートタイプがqc-coverage-region-iに要求された場合) qc-coverage-region-i_callability.bed(GVCFモードが有効で、コール可能性またはエクソーム-コール可能性レポートタイプが要求された場合)
はい	いいえ	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(GVCFモードが有効な場合)

--vc-target-bedを指定しますか? はい/いいえ	--qc-coverage-region-iを指定しますか (iは1、2、3のいずれか)? はい/いいえ	期待される出力ファイル
はい	はい	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(GVCFモードが有効な場合) ユーザーが指定したカバレッジ領域ごとに: 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(full_resレポートタイプがqc-coverage-region-iに要求された場合) qc-coverage-region-i_cov_report.bed(cov_reportレポートタイプがqc-coverage-region-iに要求された場合) qc-coverage-region-i_callability.bed(GVCFモードが有効になっており、コール可能性またはエクソーム-コール可能性レポートタイプが要求された場合)

カバレッジメトリクスのBigWigによる圧縮

*--enable-metrics-compression*をtrueに設定すると、解像度1 bpのカバレッジメトリクス出力BEDファイル（`_full_res.bed`）がBigWigフォーマットに圧縮されます。

リードカバレッジレポート

`read_cov_report`は、タブ区切りフォーマットの`_read_cov_report.bed`ファイルを生成します。最初の5つの列は、`chrom`、`start`、`end`、`name`、および`gene_id` BEDフィールドです。次の追加の列は、同じレコード行で指定された間隔領域全体にわたって計算された統計を表します。

- `total_cvg`：カバレッジ値合計。

- `read1_cvg`：リード1カバレッジ値合計。
- `read2_cvg`：リード2カバレッジ値合計。

アライメントが複数の領域と重複している場合、そのアライメントは重複が最も大きい領域に向かってカウントします。アライメントが複数の領域と均等に重複する場合、そのアライメントは最初に交差する領域に向かってカウントします。

次の例は、`_read_cov_report.bed`ファイルの内容を示しています。

```
#chrom start end name gene_id total_cvg read1_cvg read2_cvg
chr21 10033000 10034919 48 24 24
chr21 10034919 10034920 0 0 0
chr21 10034920 10034921 0 0 0
```

GCバイアスレポート

GCバイアスレポートは、GCコンテンツとそれに関連するリードカバレッジに関するゲノム全体の情報を提供します。DRAGEN GCバイアスメトリクスはPicard実装をモデル化し、既存の内部尺度に適合させています。DRAGEN GCバイアス補正モジュールは、ターゲットカウントステージに従って、これらのバイアスを補正しようとします。詳細については、141 ページの 「GCバイアス補正」 を参照してください。

GCバイアスメトリクスは、次のように計算されます。

1. デコイおよび代替コンティグを除く、リファレンスゲノム中の全染色体について、100 bp幅の塩基ごとのローリングウィンドウを用いてGCコンテンツを計算します。リファレンスにおいて4つを超えるマスクされた（N）塩基を含むウィンドウは破棄されます。
2. PF以外のリード、重複リード、二次リード、および補足リードを除く、各ウィンドウの平均カバレッジを計算します。
3. ゲノム全体の平均カバレッジを計算します。
4. GCコンテンツの割合（%）に基づいて有効なウィンドウをグループ化します。個々の割合と5つの20%範囲の両方で要約します。
5. binの平均カバレッジをゲノム全体のグローバル平均カバレッジで除することによって、各グループのノーマライズしたカバレッジを計算します。値が1.0未満の場合は、指定されたGCパーセントまたは範囲でのカバレッジが予想より低いことを示します。大きなGC値ではカバレッジが1.0から大きくずれることが結果として予測されます。
6. ドロップアウトメトリクスは、各GCが50%以下で、ATおよびGCドロップアウトが50%を上回る場合に、正の値すべての合計（GC XにおけるGC Xパーセンテージのアライメントリードでのウィンドウのパーセンテージ）として計算します。

初期設定では、GCバイアスメトリクスレポートは計算されません。GCバイアス計算を有効にするには、コマンドラインオプション*--gc-metrics-enable*を入力します。次に、コマンドの例を示します：

```
$ dragen -b <BAM file> -r <reference genome> --gc-metrics-enable=true
```

GCメトリクスレポートは、`gc_metrics.csv`ファイルを生成します。ファイルの構造は次のとおりです。

```
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>
```

GCバイアスレポートには、次のコマンドラインオプションも含まれていますが、推奨されません。

- *--gc-metrics-window-size*：初期設定のローリングウィンドウサイズ100 bpを上書きします。
- *--gc-metrics-num-bins*：要約binの数を上書きします。

体細胞メトリクスレポート

体細胞Tumor-Normalモードでは、DRAGENは腫瘍サンプルと正常サンプルについて別々のレポートを作成します。各レポートには、サンプルタイプに従ってラベルが付けられます。腫瘍サンプルレポートにはファイル名の末尾に`tumor`が含まれており、正常サンプルレポートにはファイル名の末尾に`normal`が含まれています。腫瘍サンプルと正常サンプルの両方の結果を1つのファイルに含めるには、*--vc-enable-separate-t-n-metrics*オプションをfalseに設定します。その後、DRAGENは代わりに両サンプルの総計についてレポートします。

DRAGENは以下のレポートを生成します：

- coverage_metrics
- contig_mean_cov
- fine_hist
- hist
- overall_mean_cov
- ploidy
- cov_report（要求された場合）
- full_res（要求された場合）
- gc_metrics（要求された場合）

ターゲットBEDがランに含まれているか、オプション*--qc-region-coverage-region-i*（iは1、2、または3）が含まれている場合、対応するカバレッジレポートも腫瘍ファイルと通常ファイルに分割されます。

体細胞Tumor-onlyのモードでは、レポートは生殖細胞系列モードのレポートと同一であり、ファイル名にtumorまたはnormalは含まれません。

体細胞コール可能領域レポート

体細胞モードでは、DRAGENは体細胞コール可能領域レポートをBEDファイルとして自動生成します。体細胞コール可能領域レポートには、腫瘍カバレッジが少なくとも腫瘍閾値と同程度であり、（該当する場合は）正常カバレッジが少なくとも正常閾値と同程度である全ての領域を含みます。腫瘍サンプルのみが提供された場合、レポートには腫瘍カバレッジが腫瘍閾値以上である全ての領域が含まれます。BED出力ファイルの各行は次のようにフォーマットされます。

```
chromosome region_start region_end
```

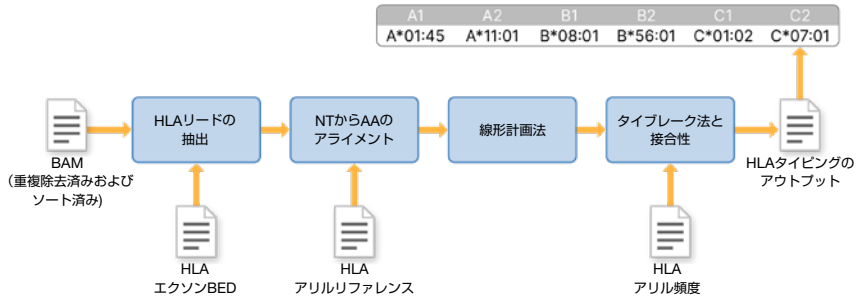
閾値は、*--vc-callability-tumor-thresh*オプションまたは*--vc-callability-normal-thresh*オプションを使用して指定できます。腫瘍閾値の初期設定値は15です。正常閾値の初期設定値は5です。各オプションの詳細については、105 ページの 「体細胞モードオプション」 を参照してください。

ターゲットベッドまたは*--qc-region-coverage-region-i*（iは1、2、または3）オプションがランに含まれている場合、DRAGENは、全ゲノム体細胞コール可能領域BEDファイルに加えて、対応する体細胞コール可能領域BEDファイルを生成します。

DRAGEN HLA Caller

DRAGENには、4桁の解像度でHLAクラスIアリルをコールするための専用のヒト白血球抗原の遺伝型が含まれています。この解像度は、HLA命名法では第2区域解像度とも呼ばれ、HLAタンパク質を判断します。HLA命名法に関する詳細については、Nomenclature for factors of the HLA system[1]を参照してください。

`--enable-hla flag`をtrueに設定することで、HLAタイピングを有効にすることもできます。以下の画像は、DRAGEN HLA Callerの概要を表しています。



DRAGEN HLA Callerは4つの基本ステップを実行します。

1. 使用するゲノムリファレンスのバージョンに応じて、指定HLA遺伝子座のリードを抽出します。必要な入力ファイルに関する詳細については、242 ページの 「HLA領域のBED入力ファイル」 を参照してください。
2. 指定HLAアリルのリファレンスシーケンスに対してアライメントを実施するには、アミノ酸を使用します。必要なリファレンス入力ファイルに関する詳細については、243 ページの 「HLAアリルのリファレンス入力ファイル」 を参照してください。

3. 線形計画法（ILP）を使って、可能性があるHLAアリル候補の短いリストを同定します。
4. タイブレーク法と接合性チェックにより、最も妥当なHLA遺伝型を同定します。指定集団レベルでのHLAアリル頻度ファイルが必要です。詳細については、244 ページの 「HLAアリル頻度の入力ファイル」 を参照してください。

以下のコマンド例により、後述のインプットの初期設定を含む、HLAタイピングが可能になります。

```
dragen \
--enable-hla=true \
--hla-bed-file=hla_exons_grch38.bed \
--hla-reference-file=hla_classI_ref_freq.fasta \
--hla-allele-frequency-file=hla_classI_allele_frequency.csv \
--hla-tiebreaker-threshold 0.97 \
--hla-zygosity-threshold 0.15 \
--output-directory={output_directory} \
--output-file-prefix={prefix} \
--enable-map-align=true \
--RGID=read_group_ID \
--RGSM=read_group_sample \
--ref-dir={reference_directory} \
--enable-map-align-output=true \
--enable-sort=true \
--enable-duplicate-marking=true \
-1 {fq1} \
-2 {fq2} \
```

HLA領域のBED入力ファイル

HLA領域のBED入力ファイルを使用して、領域を指定し、HLAリードを抽出します。HLA領域のBEDファイルを指定するには、`--hla-bed-file`を使用します。DRAGEN HLA Callerは、BEDファイル内の領域に対する入力ファイルを解析してから、HLAアリルリファレンスとアライメントするためにリードを適切に抽出します。

DRAGENが自動検出可能なヒトリファレンスゲノムを使用する場合、HLA Callerは対応するBEDファイルを選択します。自動検出が不可能なカスタマイズしたリファレンスを使用し、BEDファイルを指定しない場合、HLA Callerは警告を示し、代わりにすべてのマッピングされたリードを使用します。

以下は有効なBEDファイルの例です。

```
chr6 29942554 29942627 hla_a 1 +
chr6 29942757 29943027 hla_a 2 +
chr6 29943268 29943544 hla_a 3 +
chr6 29944122 29944398 hla_a 4 +
chr6 29944500 29944617 hla_a 5 +
```

```
chr6 29945059 29945092 hla_a 6 +
chr6 29945234 29945282 hla_a 7 +
chr6 31357086 31357159 hla_b 1 -
chr6 31356688 31356958 hla_b 2 -
chr6 31356167 31356443 hla_b 3 -
chr6 31355317 31355593 hla_b 4 -
chr6 31355107 31355224 hla_b 5 -
chr6 31354633 31354666 hla_b 6 -
chr6 31354483 31354527 hla_b 7 -
chr6 31271999 31272072 hla_c 1 -
chr6 31271599 31271869 hla_c 2 -
chr6 31271073 31271349 hla_c 3 -
chr6 31270210 31270486 hla_c 4 -
chr6 31269966 31270086 hla_c 5 -
chr6 31269493 31269526 hla_c 6 -
chr6 31269338 31269386 hla_c 7 -
```

HLAアリルのリファレンス入力ファイル

HLAアリルのリファレンス入力ファイルを使って、アライメントに対応するリファレンスアリルを指定することができます。HLAアリルのリファレンスファイルを指定するには、`--hla-reference-file`コマンドラインオプションを使用します。インプットHLAリファレンスファイルは、FASTA形式である必要があり、エクソンに分かれたタンパク質配列を含んでいます。

`--hla-reference-file`を指定しない場合は、DRAGENは`/opt/edico/config/`の`hla_classI_ref_freq.fasta`を使用します。リファレンスHLAシーケンスは、IMGT/HLAデータベースから取得します。

以下は有効なリファレンスファイルの例です。

```
>A*01:01-E1
MAVMAPRTLLLLLSGALALTQTWAG
>A*01:01-E2
SHSMRYFFTSVSRPGRGEPRFIAVGYVDDTQFVRFDSDAASQKMEPRAPWIEQEGPEYWDQETRNMKAHSQTD
RANLGTLRGYYNQSEDG
>A*01:01-E3
SHTIQIMYGCDVGPDGRFLRGYRQDAYDGKDYIALNEDLRSWTAADMAAQITKRKWEAVHAAEQRRVYLEGRC
VDGLRRYLENGKETLQRTD
>A*01:01-E4
PPKTHMTHHPISDHEATLRCWALGFYPAEITLTWQRDGEDQTQDTELVETRPAGDGTFQKWAAVVVPSGEEQR
YTCHVQHEGLPKPLTLRWE
>A*01:01-E5
LSSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSD
```

```
>A*01:01-E6
RKGGSYTQAAS
>A*01:01-E7
SDSAQGSDVSLTACKV
>A*01:03-E1
MAVMAPRTLLLLLSGALALTQTWAG
>A*01:03-E2
SHSMRYFFTSVSRPGRGEPRFIAVGYVDDTQFVRFDSDAASQKMEPRAPWIEQEGPEYWDQETRNMKAHSQTD
RANLGTLRGYYNQSEDG
>A*01:03-E3
SHTIQMMYGCDVGPDGRFLRGYRQDAYDGKDYIALNEDLRSWTAADMAAQITKRKWEAVHAAEQRRVYLEGRC
VDGLRRYLENGKETLQRTD
>A*01:03-E4
PPKTHMTHHPISDHEATLRCWALGFYPAEITLTWQRDGEDQTQDTELVETRPAGDGTFQKWAAVVVPSGEEQR
YTCHVQHEGLPKPLTLRWE
>A*01:03-E5
LSSQPTIPIVGIIAGLVLLGAVITGAVVAAVMWRRKSSD
>A*01:03-E6
RKGGSYTQAAS
>A*01:03-E7
...
```

HLAアリル頻度の入力ファイル

複数のHLAアリルが同一の結果または同様の結果を生成する場合、集団レベルのHLAアリル頻度ファイルを使用して、タイブレークすることができます。HLAアリル頻度のファイルを指定するには、`--hla-allele-frequency-file`コマンドラインオプションを使用します。HLAアリル頻度の入力ファイルは、CSV形式であり、集団のHLAアリルおよび発生頻度を含む必要があります。

`--hla-allele-frequency-file`を指定しない場合は、DRAGENは自動的に`/opt/edico/config/`の`hla_classI_ allele_frequency.csv`を使用します。集団レベルのアリル頻度は、Allele Frequency Net Databaseから取得することができます。

以下は有効なアリル頻度ファイルの例です：

```
A*01:01,305
A*01:02,140
A*01:03,100
A*01:04N,13
A*01:06,58
A*01:07,17
```

```
A*01:08,14
A*01:09,25
...
```

HLAのオプション

以下のオプションを使用して、HLA Callerを設定することができます。

- `--hla-tiebreaker-threshold`：複数のアリルに対して、同様の数のアライメントしたリードがあり、最良のアリルを示す明確なインジケーターがない場合、このアリルをタイ（同順位）と考えます。HLA Callerは、集団アリル頻度に基づきタイブレークする候補セットに、この同順位のアリルを入れます。（トップヒットにノーマライズした）アライメントしたリードの指定断片以上のアリルがある場合、アリルはタイブレークする候補セットに含まれます。初期設定値は0.97です。
- `--hla-zygosity-threshold`：メジャーアリルのリードカウントの断片に比べ、所定の座位のマイナーアリルのマップされたリードが少ない場合、HLA Callerが所定のHLA-I遺伝子のホモ接合性を推測します。このオプションを使って、その断片の値を指定します。初期設定値は0.15です。
- `--hla-min-reads`：十分なカバレッジを確保し、HLAタイピングを実施するために、HLAアリルにアライメントするリードの最小数を指定します。初期設定値は1000で、WESサンプルに対して推奨されます。カバレッジが低いサンプルを使用する場合、低い閾値を使用することができます。

HLA出力ファイル

DRAGEN HLA Callerが、6つのクラスIアリルでHLAタイピングの結果を生成します。主な出力ファイルは、`<prefix>.hla.tsv`と呼ばれるか、`<prefix>.hla.normal.tsv`および`<prefix>.hla.tumor.tsv`と呼ばれます。ファイルには、6つのアリルそれぞれに対する1列のヘッダー行、4桁の解像度での各アリルのHLAタイプの本文行が含まれます。

以下は出力ファイルの例です。

```
A1 A2 B1 B2 C1 C2
A*26:01 A*29:02 B*44:02 B*44:03 C*05:01 C*16:01
```

HLA Callerは以下の追加のHLAファイルを生成します。このファイルを使って、HLAタイピングの中間ステップを評価することができます。

- `<prefix>.freq_tiebreaking_candidates.tsv`：タイブレークプロセスの候補アリルを含んでいます。このファイルは、除外リードの数と候補アリルの集団頻度を含みます。
- `<prefix>.hla_metrics.csv`：HLA領域から抽出されたリード数、ILP選択アリルセット、ILP選択アリルセットにより解釈されるリード数、接合性を決定する除外リード数を含みます。

制限

- リード抽出ステップによって精度が低下するため、alt-awareリファレンスの使用は推奨されません。マスクしたalt-awareリファレンスを使ってHLAタイピングの精度を高めることができます。この場合、プライマリーアセンブルと類似性が高い別のハプロタイプの領域がNでマスクされます。
- DRAGENはクラスI HLA遺伝子のHLAタイピングにのみ対応します。

例

HLA Callerは、HLA領域BEDファイル、HLAアリルのリファレンスファイル、HLA頻度ファイルと同様に、FASTQまたはBAM形式の標準入力ファイルに対応します。指定しない場合は、HLA Callerは`src/config/hla`のファイルを使用します。

以下のコマンドラインの例は、FASTQファイルのインプットと初期設定のオプションの使用を示しています。

```
dragen \
--enable-hla=true \
--enable-map-align=true \
--enable-sort=true \
--enable-duplicate-marking=true \
--output-directory={output_directory} \
--output-file-prefix={prefix} \
--ref-dir={reference_directory} \
--RGID={read_group_ID} \
--RGSM={read_group_sample} \
-1 {fq1} \
-2 {fq2} \
```

以下のコマンドラインの例は、BAMファイルのインプットと初期設定のオプションの使用を示しています。

```
dragen \
--enable-hla=true \
--enable-map-align=true \
--enable-sort=true \
--enable-duplicate-marking=true \
--output-directory={output_directory} \
--output-file-prefix={prefix} \
--bam-input {bam} \
--ref-dir={reference_directory} \
```

以下のコマンドラインの例は、Tumor-Normalペアファイルのインプットと初期設定のオプションの使用を示しています。

```
dragen \
--enable-hla=true \
--enable-map-align=true \
--enable-sort=true \
--output-directory={output_directory} \
--output-file-prefix={prefix} \
--ref-dir={reference_directory} \
--tumor-fastq1 {tumor_fq1} \
--tumor-fastq2 {tumor_fq2} \
--RGID-tumor={tumor_group_ID} \
--RGSM-tumor={tumor_group_sample} \
-1 {normal_fq1} \
-2 {normal_fq2} \
--RGID={normal_group_ID} \
--RGSM={normal_group_sample} \
```

[1]Marsh SG, et al. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*. 2010 75:291-455.

バイオマーカー

腫瘍変異負荷

Tumor-onlyまたはTumor-NormalモードでSNVコーラーを実行すると、腫瘍変異負荷（TMB）を算出できます。

DRAGENは以下のステップでTMBを算出します：

1. DRAGENのTumor-Normalモードで、バリアントコールを生成します。
2. 適格領域を同定します。指定された最小限必要なカバレッジを満たす領域のみが適格となります。コマンドラインを使って、対象カスタム領域を指定することができます。
3. 以下のバリアントをフィルタリングします。
   - FAILバリアント
   - ミトコンドリアバリアント
   - MNV
   - 最小深度（DP）の閾値を満たさないバリアント。`--vc-callability-tumor-thresh`コマンドラインオプションを使用して、閾値を指定します。
   - バリアントアリルの閾値を満たさないバリアント。`--tmb-vaf-threshold`コマンドラインオプションを使用して、閾値を指定します。

- 適格領域外のバリアント。
- 生殖細胞系列バリアント。以下の基準のうちいずれかを満たす場合、生殖細胞系列として判断されるバリアント。
  - VAFが0.9を超えるサンプルVCFのバリアント。
  - 1000 GenomeまたはgnomADデータベースのいずれかで見られる、集団アリルカウントが10以上のバリアント。
  - 5つを超える周囲のバリアントが生殖細胞系列である場合。
- 腫瘍ドライバー変異。集団アリルカウントが50以上のバリアントが腫瘍ドライバー変異として扱われます。`tmb-cosmic-count-threshold`コマンドラインオプションを使用して、COSMICドライバーの閾値を指定することができます。

4. 非同義バリアントを同定します。非同義イベントのみが含まれます。非同義バリアントの結果は以下のとおりです：
   - feature_elongation, feature_truncation, frameshift_variant
   - incomplete_terminal_codon_variant, inframe_deletion, inframe_insertion
   - missense_variant, protein_altering_variant, splice_acceptor_variant
   - splice_donor_variant, start_lost, stop_gained, stop_lost, transcript_truncation
5. 以下の方程式を利用して、TMBと非同義TMBを算出します。
   - TMB= フィルタリング済みバリアント/適格領域
   - 非同義TMB=フィルタリング済み非同義バリアント/適格領域

コマンドラインオプション

以下のコマンドを使ってTMBを算出します。

コマンドラインオプション	説明
--enable-tmb true	TMBを有効にします。設定すると、スモールバリアントコーラー、Illumina Annotation Engine、関連するコール可能性レポートが有効になります。
--qc-coverage-region-1	使用するコーディング領域を指定します。
-vc-callability-tumor-thresh	コーディング領域の最小カバレッジを指定します。閾値を満たさないコーディング領域は最終的なTMB算出には使用しません。
--qc-coverage-tag-1	`--qc-coverage-tag-1`を`tmb`に設定して、DRAGEN TMBを実行します。
--qc-coverage-reports-1	`--qc-coverage-reports-1`を`callability`に設定して、DRAGEN TMBを実行します。

コマンドラインオプション	説明
--enable-variant-annotation=true --variant-annotation-assembly --variant-annotation-data	Illumina Annotation Engineを有効にします。適切なアセンブリの選択およびリファレンスファイルのダウンロードの詳細については、331 ページの「Illumina Annotation Engine」を参照してください。
（オプション）--tmb-vaf-threshold	バリアントの最小VAF閾値を指定します。閾値を満たさないバリアントはフィルタリングで除外されます。初期設定値は0.05です。
（オプション）--tmb-db-threshold	生殖細胞系列バリアントと考えられる、gnomADまたは1000 Genomeのアリルの最小アリルカウント（観測値の合計）を指定します。同じ位置および同じアリルを持つバリアントコールはTMB算出から無視されます。初期設定値は10です。

出力

TMB値は`<output prefix>.tmb.metrics.csv`で出力されます。ファイルフォーマットは、その他のメトリクスCSVファイルと同様に、以下のCSV列の規則を用います。

メトリクス	説明
Eligible Region (Mbp)	最小カバレッジ閾値を満たす指定カスタム領域。
Filtered Variant Count	VAF、バリアントタイプ、生殖細胞系列のフィルタリング後に残ったバリアント。
Filtered Nonsyn Variant Count	非同義イベントのみを含むようにさらにフィルタリングした、フィルタリング済みのバリアント。
TMB	適格領域でフィルタリング済みのバリアント。
Nonsyn TMB	適格領域でフィルタリング済みの非同義バリアント。

マイクロサテライト不安定性

マイクロサテライトは、5～50回繰り返す短いDNAモチーフのゲノム領域であり、高い変異率に関連しています。マイクロサテライト不安定性（MSI）は、DNAミスマッチ修復パスウェイでの機能低下により生じ、複数のがん種での免疫療法の反応性を予測する重要なバイオマーカーとして使用できます。

マイクロサテライト不安定性（MSI）は、Tumor-NormalモードでDNA SNVコーラーを実行する際に算出できます。

マイクロサテライト領域は正確にマッピングすることが困難ですが、DRAGENを用いることで*in silico*でMSIを推測することができます。

`--msi-command tumor-normal`フラグを含むことで、MSI解析を有効にすることができます。このコマンドにより、以下のステップが実行されます。

1. 指定したマイクロサテライト領域ごとのリードデータの腫瘍カウントと正常カウントを一覧にします。
2. 不十分なカバレッジの部位をフィルタリングで除外します。
3. カイ二乗検定を実行することで、不安定な部位を判断します。不安定な部位では、腫瘍と正常の間で大きく変化する反復長の分布がみられます。
4. サンプルごとの不安定な部位の割合を示すMSIスコアのレポートを生成します。

コマンドラインオプション

以下のコマンドを使ってMSIを算出します：

コマンドラインオプション	説明
--msi-command tumor-normal	MSIを有効にします。
--msi-microsatellites-file	マイクロサテライトを含むファイルを指定します。このファイルは、MSI-sensorを使ってマイクロサテライトのゲノムをスキャニングして、生成することができます。DRAGENは10 bp以上のホモポリマーで検証を行っています。
--msi-coverage-threshold	マイクロサテライトの最小スパニングリードカバレッジを指定します。指定閾値を満たさないマイクロサテライトは解析に含まれません。DRAGENでは、指定閾値として60を使用することを推奨します。

以下がマイクロサテライトファイルの例です：

```
#chromosome location repeat_unit_length repeat_unit_binary repeat_times
left_flank_binary right_flank_binary repeat_unit_bases left_flank_bases
right_flank_bases
chr1 985443 1 2 15 676 992 G GGGCA TTGAA
chr1 7980985 1 0 10 231 1020 A ATGCT TTTTA
chr1 8022800 1 3 19 13 41 T AAATC AAGGC
chr1 8029500 1 2 10 39 0 G AAGCT AAAAA
chr1 9146447 1 3 15 887 248 T TCTCT ATTGA
chr1 9767837 1 3 12 704 195 T GTAAA ATAAT
```

デフォルト入力ファイルを含む、MSI解析を有効にするコマンドの例は、以下のとおりです。

```
dragen \
--msi-command tumor-normal \
```

```
--msi-coverage-threshold 60 \
--msi-microsatellites-file msi_file \
--output-directory={output_directory} \
--output-file-prefix={prefix} \
--enable-map-align=true \
--RGID=read_group_ID \
--RGSM=read_group_sample \
--ref-dir={reference_directory} \
--enable-map-align-output=true \
--enable-sort=true \
--enable-duplicate-marking=true \
-1 {fq1} \
-2 {fq2} \
```

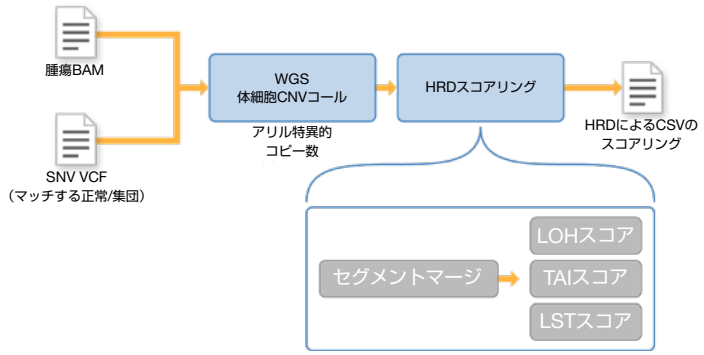
MSI出力

MSI値は`<output prefix>.microsat_output.json`で出力されます。

メトリクス	カウント
PercentageUnstableSites	18.5
TotalMicrosatelliteSitesAssessed	1581
TotalMicrosatelliteSitesUnstable	369

Homologous Recombination Deficiency

DRAGEN Homologous Recombination Deficiency（HRD）スコアリングでは、アリル特異的なコピー数のコールがVCFフォーマットで取り込まれるか、体細胞コピー数コーラーから直接ストリームされます。その後に、DRAGEN HRDはヘテロ接合性欠失（LOH）、テロメアアリル不均衡（TAI）、および大規模な状態遷移（LST）のスコアを算出します。この3つのスコアは`.hrdscore.csv`ファイルに出力されます。WGS体細胞CNVコールの結果をインプットする場合にのみ、DRAGEN HRDを使用できます。



コマンドラインオプション

以下のコマンドラインオプションを使ってHRDスコアリングを実行します。HRDスコアリングは、体細胞CNVコールとともに実行するか、体細胞CNVコールの結果を使用した後に実行可能です。

体細胞CNVコールとともにHRDスコアリングを実行するには、以下のオプションを使用します：

- *--enable-hrd*：HRDスコアリングを有効にして、ゲノムの不安定性を定量するには、trueに設定します。
- *--enable-cnv*：CNVコールを有効にして、HRDスコアリングとともに実行するには、trueに設定します。

体細胞CNVコールの後にHRDスコアリングを実行するには、以下のオプションを使用します：

- *--enable-hrd*：HRDスコアリングを有効にして、ゲノムの不安定性を定量するには、trueに設定します。
- *--hrd-input-ascn*：体細胞CNVコールの後にHRDスコアリングを実行する場合、アリル特異的なコピー数のファイル（`*cnv.vcf.gz`）を指定します。
- *--hrd-input-tn*：体細胞CNVコールの後にHRDスコアリングを実行する場合、腫瘍でノーマライズしたビンカウントファイル（`*.tn.tsv.gz`)を指定します。

HRD出力

以下のメトリクスは`.hrdscore.csv`出力ファイルに含まれています。以下は出力ファイルの例です。

サンプル	LOH_Score	TAI_Score	LST_Score	HRD_Score
サンプル	16	17	28	61

ダウンサンプリング

DRAGENはダウンサンプリングを使って、通常のアライメントの出力とは別のリードのランダムサブセットを取っておくことができます。ダウンサンプリングを使って、サンプル間または複製物間を比較するためのデータセットを生成できます。DRAGENは、ハードウェアアクセラレーションによるトリミングまたは機能のフィルタリング後に、リードをサンプリングします。これにより、DRAGENは解析リードテストのデータセットを迅速に作成できます。

ダウンサンプリングを有効にするには、`--enable-down-sampler`コマンドラインオプションをtrueに設定します。

DRAGENホストソフトウェアと互換性がある有効なシーケンスデータフォーマットを使用できます。互換性があるインプットオプションの詳細については、53 ページの 「入力オプション」 を参照してください。

DRAGENのダウンサンプリングにより、取っておいたデータのサブセットがFASTQフォーマットで出力されます。インプットがペアエンドである場合、DRAGENはサブサンプルデータを含む2つのFASTQファイルを出力します。インプットがペアエンドでない場合、DRAGENは2つのFASTQファイルを出力します。

コマンドラインの設定

ダウンサンプリングのコマンドラインオプションを有効にするだけでなく、ダウンサンプリングするリード量を設定する必要があります。リード量を設定するには、`--down-sampler-reads`か`--down-sampler-coverage`を使用します。

カバレッジレベルが指定済みの場合、`--ref-dir`を使ってゲノムを指定するか、`--down-sampler-genome-size`を使ってゲノムサイズを手動で指定する必要があります。リードおよびカバレッジの両方の制限値を指定する場合、DRAGENは両方の量の制限値を適用し、結果の値が小さい方を維持します。

オプション	説明
--enable-down-sampler	ダウンサンプリングを有効にするには、`true`を設定します。初期設定値はfalseです。 有効にする場合、`down-sampler-reads`または`--down-sampler-coverage`を設定する必要があります。
-down-sampler-num-threads	ダウンサンプリングリードに使用するスレッド数を指定します。 初期設定値は8です。
--down-sampler-random-seed	ダウンサンプリングリードの乱数シードを設定します。初期設定値は42です。
--down-sampler-genome-size	ダウンサンプリングカバレッジのターゲットゲノムサイズを設定します。初期設定値は0です。 `--down-sampler-genome-size`オプションは`--ref-dir`オプションとともに使用できません。
--down-sampler-reads	ダウンサンプリングのリードのターゲット数を指定します。初期設定値は0です。
--down-sampler-coverage	ダウンサンプリングのターゲットゲノムカバレッジを設定します。初期設定値は0です。 有効にする場合、`-ref-dir`または`--down-samplergenome-size`を設定する必要があります。

Virtual Long Read Detection

DRAGEN Virtual Long Read Detection（VLRD）は別のバリアントコーラーで、ゲノムの相同領域または類似領域の処理に特化した、より正確なバリアントコーラーです。従来のバリアントコーラーはマッパーやアライナーを利用して、どのリードが所定の位置から発生した可能性が高いかを判断します。また、所定の位置に直接隣接していないその他の領域とは無関係に、所定の位置の基本配列を検出します。従来のバリアントコールは、単一のリード長内で対象の領域がゲノムのその他の領域と類似していない場合に、うまく機能します。

ただし、ヒトゲノムの大きな断片はこの基準に当てはまりません。ゲノムの多くの領域には、ほぼ同一のコピーがどこかに存在しており、その結果、リードにおける真のソース位置の不確実性がかなり高くなる場合があります。低い信頼度でリードの集団がマッピングされる場合、一般的なバリアントコーラーでは、例えこのリードに有益な情報が含まれていたとしても、このリードが無視される可能性があります。リードがミスマップされる場合（一次アライメントがリードの真のソースではない場合）、検出エラーが生じる恐れがあります。特にショートリードのシーケンス技術ではこういった問題が生じる傾向があります。ロングリードのシーケンスはこういった問題を軽減できますが、通常、コストやエラー率が非常に高いといった短所があります。

DRAGEN VLRDはショートリードのデータから得た観点に基づき、ゲノムの反復が示す複雑性に取り組んでいます。VLRDは、各領域を切り離して考えず、リード集団が発生した可能性があるあらゆる位置を検討し、利用可能な情報すべてを一緒に用いて基本配列を検出しようとします。

DRAGEN VLRDの実行

VLRDは、DRAGENのバリアントコーラーと同じように、FASTQかソート済みのBAMファイルをインプットとして取り込み、出力VCFファイルを生成します。VLRDは一連の相同領域（2つの相同領域だけで構成）に対応しています。

VLRDは初期設定では有効化されていません。VLRDを実行するには、*--enable-vlrd*オプションをtrueに設定します。以下はVLRDを実行するためのDRAGENコマンドの例です。

```
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
```

更新したVLRDマップ/アライメント出力

DRAGEN VLRDは、通常のDRAGENマップ/アライメント出力に加えて、再マップしたBAM/SAMファイルを出力することができます。再マップしたBAM/SAMファイルの出力を有効にするには、*--enable-vlrd-map-align-output*オプションをtrueに設定します。このオプションの初期設定はfalseです。

追加のVLRDマップ/アライメント出力には、VLRDが処理した領域にマップしたリードが含まれます。

VLRDはリードのアライメントを更新するために、すべての相同領域から得られる情報を一緒に検討します。更新したVLRDマップ/アライメント出力は、相同領域が関与する解析の集積に主に使用されます。

VLRD設定

以下はDRAGENホストソフトウェアのVLRD特有のオプションです。

オプション	説明
--enable-vlrd	trueに設定すると、VLRDはDRAGENパイプラインに対して有効になります。

オプション	説明

--vc-target-bed	インプットBEDファイルを指定します。DRAGENはVLRDで処理する相同領域を指定するインプットターゲットBEDファイルを必要とします。
	インプットBEDファイルは相同領域を正確に処理するようフォーマットされています。VLRDが処理する領域の最大長は900 bpです。
	例えば:
	`chr1 161497562 161498362 0 0`
	`chr1 161579204 161580004 0 0`
	`chr1 21750837 21751637 1 0`
	`chr1 21809355 21810155 1 1`
	• 最初の3つの列は従来のBEDファイルと類似しており、1列目は染色体の説明、2列目は領域の開始、3列目は領域の終了を示しています。
	• 4列目は相同領域のグループIDです。これにより、互いに相同の領域をグループ化します。
	• 1行目と2行目の4列目は同じ値であり、一連の相同領域として処理する必要があることを示しており、3行目と4行目の次のグループとは無関係です。設定が誤っていると、ソフトウェアは互いに相同ではない領域をグループ化し、誤ったバリアントコールが生じる可能性があります。
	• 5列目は、一方の相同領域に対して、領域が逆相補しているかを示しています。値1は、同グループの一方の領域に対して、領域が逆相補していることを示しています。
	• 4行目の5列目は1に設定します。これは、逆相補に限り、この領域が3行目の領域に対して相同であることを示しています。
	DRAGENのインストールパッケージには、VRLDの2つのBEDファイル（hg19およびhs37d5リファレンスゲノム）が含まれており、これらのファイルは`/opt/edico/examples/VLRD`にあります。このBEDファイルは、VLRDを実行するためにそのままの状態で使用するか、カスタムBEDファイルを作成するための見本として使用できます。

--enable-vlrd-map-align-output	trueに設定すると、VLRDは、VLRDが処理した領域にマップしたリードのみを含む再マップしたBAM/SAMファイルを出力します。

Unique Molecular Identifiers

DRAGENは、ユニーク分子識別子（UMI）により、全ゲノムおよびハイブリッドキャプチャーアッセイからのデータを処理できます。UMIは、増幅前にDNA断片に追加される分子タグであり、増幅した断片の元のインプットDNA分子を測定します。UMIは、ライブラリー調製、PCRエラー、またはシーケンスエラーの前の脱アミノ化のようなDNA損傷により導入されるエラーとバイアスを低減させるのに有効です。

UMIパイプラインを使用するには、入力リードファイルをペアエンドランからのファイルにする必要があります。入力は、FASTQファイルのペアまたはアライメントされている/アライメントされていないBAM入力にすることができます。

DRAGENは、以下のUMIタイプに対応しています：

- デュアル、ランダムではないUMI。例えば、TruSight Oncology（TSO）UMI ReagentsやIDT xGen Prism。
- デュアル、ランダムのUMI。例えば、Agilent SureSelect XT HS2分子バーコード（MBC）やIDT xGen Duplex Seq Adapters。
- シングルエンド、ランダムのUMI。例えば、Agilent SureSelect XT HS分子バーコード（MBC）やIDT xGenデュアルインデックスUMI Adapters。

DRAGENではUMIシーケンスを使用して、元の入力断片ごとにリードペアをグループ化し、このようなグループ、つまりファミリーごとにコンセンサスリードペアを生成します。コンセンサスでエラー率を低減させることにより、DNAサンプル中のまれで低頻度の体細胞バリアントを高精度で検出します。DRAGENは、以下のようにしてコンセンサスを生成します。

1. リードをアラインメントします。
2. UMIが一致するペアのアライメントのグループにリードをグループ化します。これらのグループは、ファミリーと呼ばれています。
3. リードファミリーごとに単一のコンセンサスリードペアを生成します。

生成されたこれらのリードは、入力リードよりクオリティスコアが高く、複数の観察を各ベースコールに組み合わせることで得られた信頼度の向上を反映しています。

UMIワークフローは、DRAGENのスモールバリアントコールおよびSVにのみ対応しています。

UMI入力

以下のフォーマットのいずれかでUMIを入力します：

- **リード名**：UMIシーケンスは、リード名（QNAME）の8番目のコロン区切りフィールドに置かれています。例えば、NDX550136:7:H2MTNBDXX:1:13302:3141:10799:AAGGATG+TCGGAGA。
- **BAMタグ**：UMIは、事前アライメント済みまたはアライメントされているBAMファイル（標準のSAM形式）のRXタグとして表されます。
- **FASTQファイル**：UMIは、リードペアと同じリード順序を使用して、3番目のFASTQファイルに置かれています。

FASTQを生成するには、リード名の最後にUMIを付加してから、DRAGEN BCL変換ツールで適切なOverrideCycles設定を指定します（307 ページの 「BCL変換」 を参照）。DRAGENは、それぞれが最大8 bpで+で区切られている2つの部分からを含むUMI、または最大15 bpの単一のUMIをサポートしています。

UMIワークフローは、ユニークセットのRGSM/RGLBに対応する1セットのリードを使用して実行する必要があります。すべてのレーンが同じRGSM/RGLBセットに対応している場合、DRAGENは複数のレーンをサポートしています。

Tumor-Normalランは2つの異なるRGSMに対応しているため、DRAGEN UMIではTumor-Normal解析をサポートしていません。Tumor-Normalランでは、tumorに対して1つのサンプル名を使用し、normalに対して1つのサンプル名を使用します。DRAGEN UMIは、1回のランで1つのサンプルをサポートしています。

入力としてBAMファイルまたはFASTQファイルのリストを使用する場合、入力に複数のサンプルが含まれている場合があります。DRAGENは、ランに含まれているのは1つのサンプルのみであるかどうか、およびサンプルで使用されているのは単一でユニークRGLBライブラリーのみであるかどうかを確認します。またDRAGENは、複数のレーンにわたって広がっていたライブラリーも受け入れます。単一のサンプルと単一のライブラリーが存在する場合、DRAGENは含まれているすべてのリードを処理します。複数のサンプルまたは複数のライブラリーが存在する場合、DRAGENではエラーが発生して解析が中止します。

UMI入力修正表

デュアル、ランダムではないUMIでは、入力として既定のUMI修正表または有効なUMIシーケンスのリストを指定できます。UMI修正表を作成するには、タブ区切りのファイルを使用して、ヘッダーを含め、以下のフィールドを追加します。

フィールド	値
UMI	UMIシーケンス。例えば、`ACGTAC`。
IsValid	UMIシーケンスが有効であるかどうかを指定します。 次のいずれかを入力します:`TRUE`または`FALSE`。
NearestCodes	最も近いUMIシーケンスのコロン区切りのリスト。 例えば、`ACGTAA:ACGTAT`。
SecondNearestCodes	2番目に近いシーケンスのコロン区切りのリスト。 例えば、`ACGGAA:ACGGAT`。

カスタマイズした修正表を指定していない場合、DRAGENは、`src/config/umi_correction_table.txt`に置かれているTruSight Oncology（TSO）UMI Reagentsの初期設定の表を使用します。または、行ごとに有効なUMIシーケンスを1つ含む、ホワイトリストに登録されているランダムではないUMIのファイルを指定できます。DRAGENは、ハミング距離が1であるUMI修正表を自動生成します。

UMIオプション

オプション	説明
--umi-library-type	異なるUMI修正のバッチオプションを設定します。異なるUMIタイプのまとめ設定を最適化する3つのバッチモードが使用できます。以下のいずれかのモードを使用します: • *random-duplex*：デュアル、ランダムのUMI。 • *random-simplex*：シングルエンド、ランダムのUMI。 • *nonrandom-duplex*：デュアル、ランダムではないUMI。このオプションを使用するには、`--umi-metrics-interval-file`を使用してターゲットマニフェストファイルを指定します。

オプション	説明
--umi-min-supporting-reads	コンセンサスリードを生成するのに必要な一致しているUMI入力リード数を指定します。対応するリードが不十分なファミリーは破棄されます。例えば、以下はFFPEおよびctDNAの推奨設定です。 • **[FFPE]** バリアントが> 1%の場合、`--umi-min-supporting-reads=1`と*--vc-enable-umi-solid*バリアントコーラーパラメーターを使用します。バリアントコーラーオプションの詳細については、バリアントの93 ページの 「バリアントコーラーオプション」 を参照してください。 • **ctDNA** バリアントが< 1%の場合、`--umi-min-supporting-reads=2`と`--vc-enable-umi-liquid`バリアントコーラーパラメーターを使用します。バリアントコーラーオプションの詳細については、93 ページの 「バリアントコーラーオプション」 を参照してください。
--umi-enable	リードのまとめを有効にするには、`--umi-enable`オプションをtrueに設定します。このオプションは`-enable-duplicate-marking`には対応していません。これは、UMIパイプラインが、重複していない最適のリードを選択するのではなく、1セットの候補入力リードからコンセンサスリードを生成するためです。`--umi-library-type`オプションを使用する場合、`--umi-enable`は必要ありません。
--umi-emit-multiplicity	コンセンサスシーケンスタイプを出力に設定します。DRAGEN UMIは、元の分子の2つのストランドからのデュプレックスシーケンスをまとめることができます。通常、デュプレックスシーケンスは合計ライブラリーの約20～60%ですが、これはライブラリーキット、入力物質、およびシーケンス深度に応じて異なります。以下のいずれかのコンセンサスシーケンスタイプを入力します: • *both*：シンプレックスおよびデュプレックスのシーケンスの両方を出力します。このオプションが初期設定です。 • *simplex*：シンプレックスシーケンスのみを出力します。 • *duplex*：デュプレックスシーケンスのみを出力します。
--umi-source	UMIシーケンスの入力タイプを指定します。有効な値は以下のとおりです:`qname`、`bamtag`、`fastq`。`--umi-source=fastq`を使用する場合、`--umi-fastq`を使用してFASTQファイルからUMIシーケンスを指定します。
--umi-correction-table	カスタマイズした修正表へのパスを入力します。初期設定では、Local Run Managerは、Illumina TruSight OncologyおよびIllumina for IDT UMI Index Anchorキット用に組み込まれたテーブルで検索修正を使用します。
--umi-nonrandom-whitelist	カスタマイズした有効なUMIシーケンスのパスを入力します。
--umi-metrics-interval-file	ターゲット領域のパスをBEDフォーマットで入力します。

ランダムではないおよびランダムのUMI修正

DRAGENは、UMIおよびアライメント位置ごとにリードをグループ化することにより、UMIを処理します。UMIにシーケンスエラーが存在する場合、DRAGENはルックアップテーブルを使用するか、またはシーケンスの類似性とリードカウントを使用して、小さなシーケンスエラーを検出して修正できます。`--umi-library-type`オプション、または`--umi-correction-scheme`オプションで値`lookup`、`random`、または`none`を使用して、修正のタイプを指定します。

ランダムではないUMIの低密度のセットでは、修正できるシーケンスおよびその修正方法を指定するルックアップテーブルを作成できます。この修正ファイルスキームは、シーケンス間のハミング/編集距離が最小であるUMIセットで最適に動作します。初期設定では、DRAGENは、Illumina TruSight OncologyおよびIllumina for IDT UMI Index Anchorキット用に組み込まれたテーブルでによる検索修正を使用します。`--umi-correction-table`オプションを使用して、修正ファイルのパスを指定します。ランダムではないUMIの異なるセットを使用する場合、対応する修正ファイルの生成については、イルミナのテクニカルサポートにご連絡ください。

ランダムのUMI修正スキームでは、DRAGENは同じ位置で観察された他のUMIと比較して、所定の位置でエラーになる可能性のあるUMIを推定する必要があります。エラーモードには、1つのミスマッチまたはライブラリー調製からのUMIジャンピングまたはホッピングアーティファクトのような、小さなUMIエラーが含まれています。DRAGENは、これを以下のようにして達成します。

- 断片アライメント位置により、リードをグループ化します。
- 各位置の小さな不明確なウィンドウ内で、ファミリーを形成している厳密なUMIシーケンスにより、最初にリードをグループ化します。
- 特定の位置でのインサートサイズ分布および明確なUMI数を通して、UMIジャンピングまたはホッピングの確度を推定します。
- 不明確なウィンドウ内で、ペアの尤度の割合を計算し、UMIシーケンスとゲノム位置の異なる2つのファミリーが元の同じ分子から導出されるかどうかを評価します。
- 尤度が閾値より低いファミリーを結合します。初期設定の閾値は1です。

デュプレックスUMIの結合

デュプレックスUMIアダプターは、二本鎖DNA断片の両方のストランドを同時にタグ付けします。元の断片の各ストランドの増幅から生成されたリードを同定できます。

DRAGENは、2つのまとめられたリードペアが、同じアライメント位置を持ち（不明確なウィンドウ内）、方向が相補的で、そのUMIがリード1とリード2から交換されている場合、それらのペアをDNAの元の同じ断片の2つのストランドのシーケンスであるとみなします。シングルエンドUMIのみが存在する場合、DRAGENは、2つのストランドからのファミリーの開始-終了位置を比較して、ペアの尤度を計算し、それらが2つの個別のファミリー由来の可能性があるか、またはデュプレックスシーケンスとして結合する必要があるかを判断します。初期設定では、DRAGENは、シンプレックスおよびデュプレックスのコンセンサスシーケンスの両方を出力します。コンセンサスシーケンスの出力タイプを変更するには、`--umi-emit-multiplicity`を使用します。

UMIコマンドの例

FASTQからのコンセンサスBAMの生成

以下は、Illumina UMIにより入力リードからコンセンサスBAMファイルを生成するためのDRAGENコマンドの例です。

```
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
```

FASTQ UMI入力の使用

他のランダムのUMIライブラリータイプで実行するには、`--umi-library-type`を`random-simplex`または`random-duplex`に変更します。

```
dragen \
-r <REF> \
-1 <FQ1> \
-2 <FQ3> \
--umi-source=fastq \
--umi-fastq <FQ2> \
--output-dir <OUTPUT> \
--output-file-prefix <
PREFIX> \
--enable-map-align true \
--enable-sort true \
--umi-library-type nonrandom-duplex \
--umi-metrics-interval-file [valid target BED file]
```

カスタマイズした修正表の使用

```
dragen \
-r <REF> \
-1 <FQ1> \
```

```
-2 <FQ2> \
--umi-correction-table <valid umi correction table> \
--output-dir <OUTPUT> \
--output-file-prefix <PREFIX> \
--enable-map-align true \
--enable-sort true \
--umi-library-type nonrandom-duplex \
--umi-metrics-interval-file <valid target BED file>
```

UMI出力

Collapsed BAM

BAM出力を有効にする場合、DRAGENは、すべてのUMIコンセンサスリードを含む`<output_prefix>.bam`を生成します。リードのQNAMEは、以下の規則に基づいて生成されます。

```
consensus_read_refID1_pos1_refID2_pos2_orientation
```

- **refID1**：リード1のリファレンスID。
- **pos1**：リード1のゲノム位置。
- **refID2**：リード2のリファレンスID。
- **pos2**：リード2のゲノム位置。
- **orientation**：リード1およびリード2の方向。方向は、以下のいずれかの値にできます。位置は、リードの最も外側のアライメント位置を参照し、ソフトクリップに対して調整されます。
  - **1**：リード1は順鎖でリード2は逆鎖です。リード1の開始位置は、リード2の終了位置以下です。
  - **2**：リード1は逆鎖でリード2は順鎖です。リード2の開始位置は、リード1の終了位置以上です。
  - **3**：リード1は順鎖でリード2は逆鎖です。リード1の開始位置は、リード2の終了位置より大きいです。
  - **4**：リード1は逆鎖でリード2は順鎖です。リード2の開始位置は、リード1の終了位置より大きいです。
  - **5**：リード1およびリード2は順鎖です。
  - **6**：リード1およびリード2は逆鎖です。

UMIメトリクス

DRAGENは、UMIのまとめの統計値を記述する`<output_prefix>.umi_metrics.csv`ファイルを出力します。このファイルには、入力リード、それらをファミリーにグループ化した方法、UMIを修正した方法、およびファミリーがコンセンサスリードを生成した方法に関する統計値が要約されています。以下のメトリクスは、アプリケーションのパイプラインのチューニング時に有効になる場合があります：

- **破棄したファミリー**：`--umi-min-supporting-reads`入力より少ないか、またはデュプレックス/シンプレックスのステータスが`--umi-emit-multiplicity`で指定したステータスとは異なるファミリーはすべて破棄されます。これらのリードは、`Reads filtered out`としてログに記録されます。ファミリーは、`Families discarded`としてログに記録されます。
- **UMI修正**：ファミリーは、さまざまな方法で組み合わせられる場合があります。このような修正数は、以下のようにレポートされます。
  - **シフトされたファミリー**：`umi-fuzzy-window-size`パラメーターで指定した距離までの断片アライメント座標を持つファミリー。`umi-fuzzy-window-size`パラメーターの初期設定は3です。
  - **文脈上修正されたファミリー**：同じ断片アライメント座標および対応するUMIを持つファミリーを結合します。
  - **デュプレックスファミリー**：アライメント座標が近くUMIが相補的なファミリーを結合します。

*--umi-metrics-interval-file*の有効なパスを指定すると、DRAGENは、指定したBEDファイル内のファミリーのみを含む別のセットのオンターゲットのUMI統計値を出力します。

観察されたUMIが可能性のあるUMIシーケンスのすべてをカバーしている範囲を解析する必要がある場合は、断片位置メトリクスごとにユニークなUMIのヒストグラムが有効になる場合があります。これはゼロベースのヒストグラムであり、ここではインデックスは特定の断片位置でユニークなUMIカウントを示しており、値はそのカウントでの位置数を表しています。

以下の表は、使用できるUMIメトリクスを示しています。

メトリクス	説明
Number of reads	リード総数。
Number of reads with valid or correctable UMIs	ルックアップテーブルに基づいて、UMIを修正できたリード数。
Number of reads in discarded families	破棄したファミリーでのリード数。ファミリーをサポートするのに十分な処理前リードが存在しない場合、ファミリーは破棄されます。
Reads with all-G UMIs filtered out	UMIシーケンスですべてGであるためにフィルタリングで除外されたリード数。
Reads filtered out	特性または破棄されたファミリーでフィルタリングして除外されたリード総数。
Reads with uncorrectable UMIs	UMIを修正できなかったリード数。

メトリクス	説明
Total number of families	シンプレックスのまとめられたリード数。
Families contextually corrected	文脈上修正されたファミリー数。UMI修正は、同じマッピング位置の他のファミリーに基づいています。
Families shifted	シフト修正されたファミリー数。シフト修正では、`umi-fuzzy-window-size`パラメーターで指定した距離までの断片アライメント座標を持つファミリーを結合します。
Families discarded	最小対応リード基準またはumi-emitタイプのシンプレックス/デュプレックスに不合格でフィルタリングで除外されたファミリー数。
Families discarded by min-support-reads	最小対応リード基準に不合格でフィルタリングで除外されたファミリー数。
Families discarded by duplex/simplex	umi-emitタイプのシンプレックス/デュプレックスに不合格でフィルタリングで除外されたファミリー数。
Families with ambiguous correction	3つ以上のUMI修正が存在する可能性があるため、UMIが修正できないファミリー数。
Duplex families	デュプレックス（両方のストランド）として結合されたファミリー数。パーセンテージの場合、分母はコンセンサスペア数です。
Consensus pairs emitted	出力BAMでまとめられたリード数。
Mean family depth	ファミリーごとの平均リード数。
Histogram of num supporting fragments	0の処理前リード、1つの処理前リード、2つの処理前リード、3つの処理前リードなどがあるファミリー数。
Number of collapsible regions	領域数。
Min collapsible region size (num reads)	最小データ入力領域のリード数。
Max collapsible region size (num reads)	最大データ入力領域のリード数。

メトリクス	説明
Mean collapsible region size (num reads)	領域ごとの平均リード数。
Collapsible region size standard deviation	領域ごとのリード数の標準偏差。
On target number of reads	UMIターゲット間隔*--umi-metrics-interval-file*と重複しているリード数。
On target number of reads with valid or correctable UMIs	エラー許容差など、ルックアップテーブルでUMIに一致し、UMIターゲット間隔と重複しているUMIがあるリード数。
On target number of reads in discarded families	UMIターゲット間隔と重複している破棄したファミリーでのリード数。
On target duplex families	UMIターゲット間隔と重複しているすべてのファミリー内でデュプレックスとして結合されるファミリー数。パーセンテージの場合、分母は目標コンセンサスペア数です。
On target mean family depth	UMIターゲット間隔と重複しているファミリーごとの平均リード数。
On target families discarded	UMIターゲット間隔と重複しており、最小対応リード基準またはumi-emitタイプのシンプレックス/デュプレックスに不合格でフィルタリングされたファミリー数。
On target families discarded by min-support-reads	UMIターゲット間隔と重複しており、最小対応リード基準に不合格でフィルタリングされたファミリー数。
On target families discarded by duplex/simplex	UMIターゲット間隔と重複しており、umi-emitタイプのシンプレックス/デュプレックスに不合格でフィルタリングされたファミリー数。
On target families with ambiguous correction	3つ以上のUMI修正が存在する可能性があるため、UMIが修正できないUMIターゲット間隔と重複しているファミリー数。

メトリクス	説明
Histogram of unique UMIs per fragment position	0のUMIシーケンス、1つのUMIシーケンス、2つのUMIシーケンスなどがある位置数。
Total families in probability model estimation	UMIジャンピング率の推定で使用されるファミリーおよび確率論的ファミリー結合で使用される断片サイズ分布の総数。
Number of potential Jumping Families	可能性のあるUMIジャンピング候補のファミリー総数と対応する比率。

マルチコーラーワークフロー

DRAGENでは、単一のワークフローで複数のツールを実行できます。

*enable-component*フラグは、コンポーネントの有効化、無効化を制御します。DRAGENは、有効化されたコンポーネントを使ってワークフローを構築し、コンポーネントの不整合を自動的に解決します。可能であれば、DRAGENはコンポーネントを並行実行します。

コンポーネントはそれぞれ、インプット設定や内部アルゴリズムパラメーター、出力ファイルやフィルター条件などの設定するための複数のオプションを持っています。詳細については、各コンポーネントのセクションを参照してください。

*output-directory*や*sample-sex*など、一部のオプションは、複数のコーラーで共有されます。

各バリアントコーラーは、VCFとメトリクス出力ファイルのセットを独自に作成します。

コンポーネントコマンドの例

```
enable-map-align
enable-sort
enable-duplicate-marking
enable-variant-caller
enable-cnv
enable-sv
```

インプット形式

DRAGENが受け入れる一般的かつ標準的なNGSインプット形式は以下のとおりです：

- FASTQ（`fastq-file1`および`fastq-file2`）
- FASTQ List（`fastq-list`）
- BAM（`bam-input`）

- CRAM（`cram-input`）

体細胞ワークフローは、腫瘍に相当する入力ファイル（`tumor-bam-input`など）を使用できます。

アライメントされていないリードから実行した場合、リードはまず、マッピング/アライメントコンポーネントを通って、アライメントを作成します。このアライメントはそのまま下流のバリアントコーラーまで進みます。あらかじめアライメントされたリードから実行した場合、DRAGENでは、マッピング/アライメントコンポーネントを使って再アライメントする、またはソース入力からの既存のアライメントを使用することができます。

マルチコーラーコマンドラインの例

ここでは、シングルコーラーのシナリオにあるコマンドラインオプションを組み合わせて、マルチコーラーワークフローを作成する際のベストプラクティスの一例を示します。この例は、以下のステップから構成されます：

- INPUTオプションを設定。
- OUTPUTオプションを設定。
- 再アライメントが望ましいかどうかに応じてMAP/ALIGNを設定。
- 用途に基づいてバリアントコーラーを設定。
- コンポーネントごとに必要なオプションを構築し、最後のコマンドラインで再利用を可能に。

```
INPUT_OPTIONS="
--ref-dir $DRAGEN_HASH_TABLE \
--fastq-file1 $FASTQ1 \
--fastq-file2 $FASTQ2 \
--RGSM $RGSM \
--RGID $RGID \
"
OUTPUT_OPTIONS="
--output-directory $OUTPUT \
--output-file-prefix $PREFIX \
"
MA_OPTIONS="
--enable-map-align true \
... <any other optional settings> \
"
CNV_OPTIONS="
--enable-cnv true \
... <any other optional settings> \
"
SNV_OPTIONS="
```

```
    --enable-variant-caller true \
    ... <any other optional settings> \
    "
    SV_OPTIONS="
    --enable-sv true \
    ... <any other optional settings> \
    "
    CMD="
    dragen \
    $INPUT_OPTIONS \
    $OUTPUT_OPTIONS \
    $MA_OPTIONS \
    $CNV_OPTIONS \
    $SNV_OPTIONS \
    $SV_OPTIONS \
    "
```

生殖細胞系列

以下の表に、サポートされているインプット形式とバリアントコーラーの一部をまとめます。この表は抜粋で、サポートされていても記載されていない機能やコーラーもあります。

生殖細胞系列	マッピング/アライメントを使ったFASTQ	BAM/CRAM	マッピング/アライメントを使ったBAM/CRAM
CNV + SNV	サポート	サポート	サポート
CNV + SV	サポート	サポート	サポート
SNV + SV	サポート	サポート	サポート
CNV + SNV + SV	サポート	サポート	サポート

体細胞

体細胞ワークフローは、腫瘍インプットと正常インプットの両方を指定します。さらに体細胞CNVコーラー用にマッチする正常SNV VCFが必要であるだけでなく、2つの入力ファイル（腫瘍とマッチする正常）が必要になる可能性があるということは、細心の注意を払う必要があることを意味します。このため、推奨されるTumor-Normalワークフローでは、まず、マッチする正常インプットを生殖細胞系列ワークフローを通して実行します。

1. マッチする正常インプットを、生殖細胞系列ワークフロー（CNV + SNV + SV + ...）を通して実行します。このワークフローにより、マッチする正常SNV VCFが生成されます。
2. 腫瘍インプットとマッチする正常インプットを、体細胞ワークフロー（CNV + SNV + SV + ...）を通して実行します。

```
    INPUT_OPTIONS="
```

```
    --ref-dir $DRAGEN_HASH_TABLE \
    --tumor-bam-input $TUMOR_BAM \
    --bam-input $NORMAL_BAM \
    "
    OUTPUT_OPTIONS="
    --output-directory $OUTPUT \
    --output-file-prefix $PREFIX \
    "
    MA_OPTIONS="
    --enable-map-align false \
    ... <any other optional settings> \
    "
    CNV_OPTIONS="
    --enable-cnv true \
    --cnv-normal-b-allele-vcf $SNV_VCF \
    ... <any other optional settings> \
    "
    SNV_OPTIONS="
    --enable-variant-caller true \
    ... <any other optional settings> \
    "
    SV_OPTIONS="
    --enable-sv true \
    ... <any other optional settings> \
    "
    CMD="
    dragen \
    $INPUT_OPTIONS \
    $OUTPUT_OPTIONS \
    $MA_OPTIONS \
    $CNV_OPTIONS \
    $SNV_OPTIONS \
    $SV_OPTIONS \
    "
```

以下の表は、Tumor-Normalモードでサポートされているさまざまな組み合わせをまとめたものです。

Tumor-normal	マッピング/アライメントを使ったFASTQ	BAM/CRAM	マッピング/アライメントを使ったBAM/CRAM
CNV + SNV	サポート	サポート	未サポート
CNV + SV	サポート	サポート	未サポート
SNV + SV	サポート	サポート	未サポート
CNV + SNV + SV	サポート	サポート	未サポート

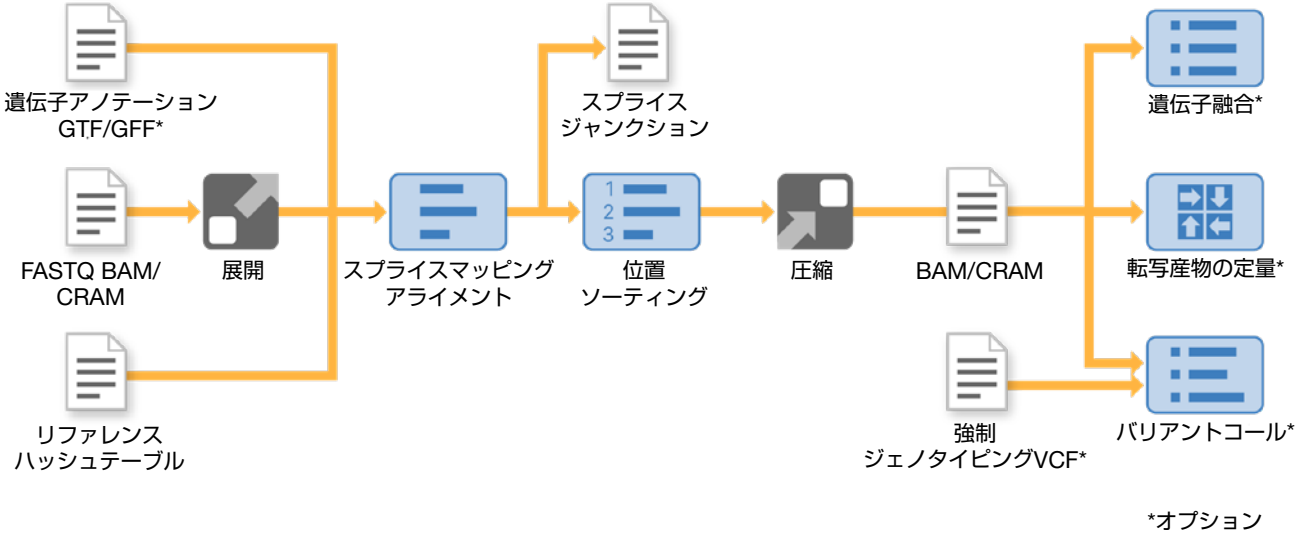
Tumor-onlyモードで実行するには、INPUTオプションからマッチする正常インプットを削除し、Tumor-onlyモードで実行するように、個々のコーラーを設定します。以下の表は、Tumor-onlyモードでサポートされているさまざまな組み合わせをまとめたものです。

Tumor-only	マッピング/アライメントを使ったFASTQ	BAM/CRAM	マッピング/アライメントを使ったBAM/CRAM
CNV + SNV	サポート	サポート	サポート
CNV + SV	サポート	サポート	サポート
SNV + SV	サポート	サポート	サポート
CNV + SNV + SV	サポート	サポート	サポート

WES解析は、そのモードがシングルコーラーモードでサポートされていて、インプット設定に不一致がない場合にサポートされます。

DRAGEN RNA Pipeline

DRAGENには、RNA-Seq（スプライシング対応）アライナーと、遺伝子発現の定量および遺伝子融合検出のためのRNA固有の解析コンポーネントがあります。



ホストソフトウェアオプションとDNAマッピングで記載された機能とオプションのほとんどがRNAアプリケーションにも適用されます。その他のRNA特異的な特長について本セクションで説明します。

入力ファイル

RNAアライメント

DRAGEN RNAパイプラインはDRAGEN RNA-Seqスプライスアライナーを使用します。RNA-Seqリードの短いシードシーケンスのマッピングは、DNAリードのマッピングと同じように行われます。さらに、マッピングされたシードに近接するスプライスジャンクション（RNA転写産物中の隣接していないエクソン同士の結合部分）を検出し、完全なリードアライメントに取り込みます。

アライメント出力

RNAモードのDRAGENを実行中に生成される出力ファイルは、DNAモードで生成されるものと同様のファイルです。RNAモードでは、スプライスされたアライメントに関連する追加の情報も産生します。スプライスジャンクションに関する詳細は、SAMアライメントレコードおよび追加ファイルのSJ.out.tabファイルの両方に存在します。

BAMタグ

出力BAMファイルはSMA仕様を満たし、下流のRNA-Seq解析ツールに対応します。次のBAMタグはスプライスされたアライメントと共に出力されます。

- **XS:A**：XSタグはイントロンのストランド方向を示します。272 ページの 「Cufflinksとの互換性」 を参照してください。
- **jM:B**：jMタグはアライメント中のすべてのジャンクションに対するイントロンモチーフを示します。次のような定義があります：
  - 0：non-canonical
  - 1：GT/AG
  - 2：CT/AC
  - 3：GC/AG
  - 4：CT/GC
  - 5：AT/AC
  - 6：GT/AT

  遺伝子アノテーションファイルをマップ / アライン段階で使用し、スプライスジャンクションがアノテーションされたジャンクションとして検出される場合、そのモチーフ値に 20 が追加されます。
- **NH:i**：標準のSAMタグは、現在のレコード中のクエリを含むレポートされたアライメント数を示します。このタグはfeatureCountsなどの下流ツールに対して使用されることがあります。
- **HI:i**：標準のSAMタグはクエリヒットインデックスを示し、このアライメントがSAMに保存されたi番目の1つであることを示す値を含みます。この値は1～NHまでの範囲があります。このタグはfeatureCountsなどの下流ツールに対して使用されることがあります。

Cufflinksとの互換性

CufflinksはXS:Aストランドタグを出力するためにスプライスされたアライメントを必要とする場合があります。このタグは、アライメントがスプライスジャンクションを含んでいる場合にSAMレコードに存在します。XS:Aストランドタグに対する値は、次のようになります：

'.' (undefined), '+' (forward strand), '-' (reverse strand), or '*' (ambiguous).

スプライスアライメントが未定義ストランドまたは対立するストランドである場合、そのアライメントは、*--no-ambig-strand*オプションを1に設定して抑制できます。

また、Cufflinksは一意にマッピングされたリードに対するMAPQが単一値であることも予想します。この値は *--rna-mapq-unique*オプションで指定されます。一意にマッピングされたリードすべてがこの値に等しいMAPQになるようにするには、*--rna-mapq-unique*を0以外の値に設定ます。

SJ.out.tab

SAM/BAMファイルに出力されたアライメントと共に、追加のSJ.out.tabファイルがタブ区切りファイルに信頼度の高いスプライスジャンクションを要約します。このファイルの列は次のようになります：

1. コンティグ名

2. スプライスジャンクションの初めの塩基（1-based）
3. スプライスジャンクションの最後の塩基（1-based）ストランド（0：未定義、1：+、2：-）
4. ストランド（0：未定義、1：+、2：-）
5. イントロンモチーフ：0：noncanonical、1：GT/AG、2：CT/AC、3：GC/AG、4：CT/GC、5：AT/AC、6：GT/AT
6. 0：未アノテーション、1：アノテーション済、遺伝子アノテーション入力ファイルが使用されている場合のみ。
7. スプライスジャンクションをまたいでユニークにマッピングしているリード数
8. スプライスジャンクションをまたいで複数にマッピングしているリード数
9. スプライスされたアライメントの最大オーバーハング値

SJ.out.tabファイル内のスプライスされたアライメントの最大オーバーハング値（列8）フィールドはアンカーアライメントのオーバーハング値です。例えば、リードがACGTACGT----------ACGTのようにスプライスされている場合、オーバーハング値は4となります。このジャンクションをまたぐすべてのリードのうち、同じスプライスジャンクションについては、最大オーバーハング値がレポートされます。最大オーバーハング値は、アンカーアライメントに基づいてそのスプライスジャンクションが正しいことを示す信頼度の指標です。

DRAGENホストソフトウェアによって生成されるSJ.out.tabファイルは2種類あり、フィルタリングなしとフィルタリングありのバージョンがあります。フィルタリングなしファイルのレコードは、出力SAM/BAMからのスプライスされたアライメントレコードすべてを統合したものです。ただし、フィルタリングありバージョンでは、次のフィルターを使用しているため、正確さについてより高い信頼度があります。

これらの条件のどれかが満たされる場合、SJ.out.tabファイルのスプライスジャンクションへのエントリーがフィルタリングで除外されます：

- SJはnoncanonicalモチーフであり、ユニークにマッピングするリードが3個未満でしか裏付けられていない。
- SJの長さが50000超であり、ユニークにマッピングするリードが2個未満でしか裏付けされていない。
- SJの長さが100000超であり、ユニークにマッピングするリードが3個未満でしか裏付けられていない。
- SJの長さが200000超であり、ユニークにマッピングするリードが4個未満でしか裏付けられていない。
- SJはnoncanonicalモチーフであり、スプライスされたアライメントの最大オーバーハング値が30未満である。
- SJはcanonicalモチーフであり、スプライスされたアライメントの最大オーバーハング値が12未満である。

フィルタリングされたSJ.out.tabファイルは下流の解析または処理ツール後の使用が推奨されます。その他に、フィルタリングされていないSJ.out.tabファイルを使用して、自身のフィルターを適用できます（例、基本的なawkコマンドでの使用）。

注：このフィルターはBAMまたはSAMファイルに存在するアライメントには適用できません。

マッピングメトリクス

RNAパイプラインは`mapping_metrics.csv`ファイル内にリードマッピングに関わるサマリーおよびリードグループあたり統計量をレポートします。メトリクスの計算はRNA中のスプライスされたアライメントを説明します。以下がメトリクスの例です。インサート長：中央値、補足（キメラ）アライメントなど。

Chimeric.out.junctionファイル

サンプルにキメラアライメントが存在する場合、補足のChimeric.out.junctionファイルも出力されます。このファイルは下流の遺伝子融合検出を実施するために使用されることがあるスプリットリードに関する情報を含みます。各ラインはキメラ状にアライメントしたリード1つを含みます。このファイルの列は次のようになります：

1. ドナー染色体。
2. ドナーのイントロンの1番目の塩基（1-based）。
3. ドナーのストランド。
4. アクセプターの染色体。
5. アクセプターのイントロンの1番目の塩基（1-based）。
6. アクセプターのストランド。
7. N/A：使用されません。その他のツールに適合するために存在します。常に1になります。
8. N/A：使用されません。その他のツールに適合するために存在します。常に*になります。
9. N/A：使用されません。その他のツールに適合するために存在します。常に*になります。
10. リード名。
11. +ストランド上の1番目のセグメントの1番目の塩基。
12. 1番目のセグメントのCIGAR。
13. 2番目のセグメントの1番目の塩基。
14. 2番目のセグメントのCIGAR。

このファイルのCIGARはSAM仕様で見られるように、標準的なCIGARオペレーションに従い、オペレーションpでエンコードされるギャップ長Lの追加を伴います。ペアエンドリードでは、2番目のメイトの配列はストランドの状態を決定する前に常に逆相補にされています。

以下は、2つのキメラ状にアライメントしたリードペアを示すエントリー例を示しており、このメイトの1つは分割されており、chr19からchr12のセグメントにマッピングしています。また、これらのエントリーと関連のある対応するSAMレコードも示されています。

```
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
```

RNAアライメントオプション

RNAスプライスアライナーのアライナーの段階では、Smith-Watermanアライメントスコアリングオプションとスプライシングスコアリングオプションを使用します。

重複マーキング

DRAGEN RNA Pipelineは重複リードを検出できます。重複リードとは、アライメント中に同一（クリッピング調節された）ポジションに両末端がマッピングするフラグメントとして定義されます。RNA-Seqデータでは、リードはライブラリー調製中のPCR重複または高発現した領域のディープカバレッジの結果として表されることがあります。
*--enable-duplicate-marking*がtrueに設定されている場合、重複フラグメントはBAMファイルにマークされ、重複リードの合計数はマッピングメトリクスとしてレポートされます。重複したというマーキングは遺伝子発現定量および遺伝子融合コールに影響しません。

リボソームRNAフィルタリング

リボソームRNA（rRNA）配列は、サンプルの種類とライブラリー調製メソッドに応じて、一部のRNA-Seqデータセットにおいて大量のリード断片の原因となることがあります。rRNAリードは下流の解析に関連しないため、DRAGEN RNAパイプラインを使用して、アライメント中にrRNAリードをフィルタリングできます。rRNAをフィルタリングすることにより、ランタイムを短縮し、ファイルサイズを削減し、ゲノム上のrRNA繰り返し座位での詳細なリードアライメントの集積を回避でき、RNA BAMファイルの下流の解析を簡単に行えます。

rRNAフィルタリングは、リファレンスハッシュテーブルに記載されたrRNA配列を含むデコイコンティグに依存します。マルチマッパーを含むデコイコンティグにマップするリードはrRNAでタグ付けされ、出力にマッピングされません。

コマンドラインオプション

以下は、rRNAフィルタリングに必要とされるコマンドラインオプションです。

- `--rrna-filter-enable=true`：rRNAフィルタリングを有効にします。rRNAフィルタリングを有効にするには、`true`に設定します。初期設定値はfalseです。
- `--rrna-filter-contig`：フィルタリングに使用するrRNA配列名を指定します。この値を指定しない場合、初期設定の`gl000220`はリファレンス自動検出機能を用いてヒトゲノムアライメントに規定されます。`gl000220`は、rRNA繰り返しの全コピーを含むhg19およびhg38ゲノムに含まれる、場所が特定されないコンティグです。その他のゲノムについては、ハッシュテーブルを作成する際にrRNAデコイコンティグを含める必要があります。

出力ファイル

すべてのrRNAフィルタリングリードは、BAMファイルにアライメントされずに残され、`ZS:Z:FLT`タグが付けられます。

フィルタリングされたrRNAリードの数および割合は、マッパーがrRNAをフィルタリングしたリードとしてレポートされます。rRNAリードは全体のマッピングされていないリードメトリクスには含められません。

MAPQスコアリング

初期設定では、RNA-Seqに対するMAPQ算出はDNA-Seqと同一です。MAPQ算出の主要な寄与因子は1番良いアライメントスコアと2番目に良いアライメントスコアの差です。したがって、アライメントスコアリングのパラメーター調節はMAPQ推定値に影響します。これらの調節は68 ページの 「Smith-Watermanアライメントスコアリング設定」 で説明しています。

*--mapq-strict-sjs*オプションはRNAに特異的であり、少なくとも1つのエクソンセグメントが高い信頼度を伴ってアライメントされている場合に適用されますが、可能性のあるスプライスジャンクションに関する曖昧さがあります。このオプションを0に設定すると、アライメントが少なくとも部分的に正しいことを示す、より高いMAPQ値が返されます。このオプションを1に設定すると、スプライスジャンクションの曖昧さを反映して、より低いMAPQ値が返されます。

Cufflinksなどの一部の下流ツールでは、ユニークにマッピングされたすべてのリードに対する固有値となるようにMAPQ値を予測します。この値は*--rna-mapq-unique*オプションで指定されます。このオプションを0以外の値に設定することにより、アライメントスコアに基づくすべてのMAPQ推定値が無効になります。その代わりに、ユニークにマッピングされたリードすべてのMAPQを*--rna-mapq-unique*の値になるように設定します。そうすることで、複数マッピングされたリードすべてがMAPQ値、int(-10*log10(1 - 1/NH)を有することになります。ここで、NH値はそのリードに対するヒット数（一次アライメントおよび二次アライメント）を表します。

遺伝子融合検出

DRAGENの遺伝子融合モジュールは、DRAGEN RNAスプライスアライナーを使用して、遺伝子融合イベントを検出します。可能性のあるブレークポイントを検出するために、補足（キメラ）アライメントに関するスプリットリード解析を実施します。推定上の融合イベントはさまざまなフィルタリングステージを経て、可能性のある偽陽性を低減します。最終結果に加えて、すべての可能性のある（フィルタリングされていない）候補が出力されるが、これらは感度を最大にするために使用することがあります。

DRAGEN遺伝子融合の実行

DRAGEN遺伝子融合モジュールは通常のRNA-Seqマッピング/アライメント作業と同時に実行できます。DRAGEN遺伝子融合モジュールを有効にするには、現在のRNA-Seqコマンドラインスクリプトで`--enable-rna-gene-fusion`をtrueに設定します。DRAGEN遺伝子融合モジュールはGTFまたはGFF形式の遺伝子アノテーションファイルを必要とします。

以下は、エンドツーエンドのRNA-Seq実験を実行するためのコマンドラインの例です。

```
/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
```

ランの終了時に、以下の例と同様の、検出された遺伝子融合イベントの概要が出力されます。

```
================================================================
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
```

遺伝子発現定量

DRAGEN RNA Pipelineは、遺伝子発現定量モジュールを搭載しており、RNA-Seqデータセットの各転写産物と遺伝子の発現を推測します。このモジュールは、まず初めに内部で、各リード（リードペア）のゲノムマッピングを一致する転写産物マッピングに置換します。次に、Expectation-Maximization（EM）アルゴリズムを使用して、観察されたリードすべてにベストマッチする転写産物の発現値を推測します。また、EMアルゴリズムは、レポートされた定量結果におけるGCバイアスのモデル化と補正を行えます。

定量モジュールを有効にするには、現在のRNA-Seqコマンドラインスクリプトで`--enable-rna-quantification`を`true`に設定します。さらに、定量するには、すべての転写産物のゲノムポジションを含む遺伝子アノテーションファイル（GTF/GFF）を提供する必要があります。`-a`または`--annotation-file`オプションを使用して、GTF/GFFを指定できます。

定量オプション

オプション	説明
--enable-rna-quantification	trueに設定した場合、RNA定量が有効になります。`--enable-rna`をtrueに設定する必要があります。

オプション	説明
--rna-quantification-library-type	RNA-Seqライブラリーの種類を指定します。以下は、使用可能な値です: • IU：ペアエンドのunstranded（どのstrandが最初に転写されたかに関する情報が保存されていない）リードのライブラリー。 • ISR：リード2が転写産物ストランドにマッチしているペアエンドのstranded（どのstrandが最初に転写されたかに関する情報が保存されている）リードのライブラリー（例、Illumina Stranded Total RNA Prep）。 • ISF：リード1が転写産物ストランドにマッチしているペアエンドのstrandedリードのライブラリー。 • U：シングルエンドのunstrandedリードライブラリー。 • SR：リードが転写産物ストランドに対して逆向き方向にあるシングルエンドのstrandedリードライブラリー（例、Illumina Stranded Total RNA Prep）。 • SF：リードが転写産物ストランドにマッチしているシングルエンドのstrandedリードのライブラリー。 • A：DRAGENがデータセット中の初めのリードペアを検証し、自動的に正しい種類のライブラリーを検出。自動検出は、初期設定値です。
--rna-quantification-gc-bias	GCバイアス補正はシーケンスカバレッジに関する転写産物の%GCの影響を推測し、発現を推測する際の影響を説明します。GCバイアス補正を無効にするには、falseに設定します。
--rna-quantification-fld-max --rna-quantification--fld-mean --rna-quantification-fld-sd	これらのオプションを使用してシングルエンドで実行するためにRNA-Seqライブラリーのインサートサイズ分布を指定します。これらのオプションはGCバイアス補正に関連します。初期設定は`250 +- 25`です。最大の許容値は`1000`です。精度を改善するには、値を修正して、ライブラリーにマッチさせます。

定量出力

転写産物の定量結果は`<outputPrefix>.quant.sf`テキストファイルにレポートされます。このファイルは各転写産物に対する結果を記載しています。tximportやDESeq2などのツールを使用して発現差のある遺伝子に対する入力として出力ファイルを使用できます。

ファイルコンテンツの例を次に示します：

```
Name                Length      EffectiveLength     TPM          NumReads
ENST00000364415.1   116         12.3238             5.2328       1
ENST00000564138.1   2775        2105.58             1.28293      41.8885
```

フィールド	説明
Name	転写産物のID。
Length	塩基対の(スプライス後の)転写産物の長さ。
EffectiveLength	インサートサイズとエッジ効果から構成される、RNA-Seqがアクセス可能な長さ。
TPM	転写産物100万個あたりの各転写産物数(TPM)。転写産物長とシーケンス深度をノーマライズした際の転写産物の発現量を表します。
NumReads	転写産物からの推定リード数。この値はノーマライズされません。

遺伝子発現定量モジュールも次のファイルを出力します。含まれるメトリクスに関する情報は、280 ページの 「定量およびRNA QCメトリクス」 を参照してください。

- `<outputPrefix>.quant.genes.sf`：遺伝子レベルの定量結果を含みます。結果は、アノテーションファイル（GTF）に同一geneIDのあるすべての転写産物を一緒にして要約して生成されます。LengthおよびEffectiveLengthは、その遺伝子の個々の転写産物について発現を加重平均した値です。
- `<outputPrefix>.quant.metrics.csv`：RNA転写産物と定量に関連する要約統計量です。280 ページの 「定量およびRNA QCメトリクス」 を参照してください。
- `<outputPrefix>.quant.transcript_fragment_lengths.txt`：転写産物にマッピングされたリードの全断片長の分布。
- `<outputPrefix>.quant.transcript_coverage.txt`：転写産物に沿った5'から3'の平均カバレッジパターンを用いたカバレッジ均一性の測定値。
- `<outputPrefix>.SJ.saturation.txt`：処理されたリードの関数として観察されたユニークなスプライスジャンクション数など、ライブラリーのシーケンス飽和度の測定値。

定量およびRNA QCメトリクス

RNA定量モジュールは、遺伝子発現結果に関連するメトリクス、および転写産物レベルの解析に基づいたより一般的なRNA QCメトリクスを出力します。

メトリクスの要約は`<outputPrefix>.quant.metrics.csv`ファイルに出力されます。

メトリクス	説明
転写産物	解析に用いられた遺伝子アノテーションファイル(GTF/GFF)入力からの転写産物数。

メトリクス	説明
推定されるライブラリー方向	オリジナルの転写産物に関連するRNA-Seqリードのライブラリー方向。ライブラリー方向を自動的に検出させるか、または方向に関する情報を提供することができます。詳細については、278 ページの「定量オプション」を参照してください。
転写産物断片	1つ以上のアノテーションされた転写産物にマッピングされた断片(リードペア)数。
転写産物のカバレッジCV中央値	転写産物に沿ったカバレッジの変動係数。このメトリクスは、RNA-Seqリードカバレッジが転写産物中にどれくらい均一にあるかを評価します。
(順鎖/逆鎖)転写産物断片	順鎖または逆鎖上の転写産物にマッチするリードペア。
ストランドミスマッチ断片	strandedタイプのライブラリーを使用している場合、転写産物の予想されるストランドにマッチしないリードペア。
方向性によるフィルタリング断片	転写産物(アンチセンス)の予想される方向にマッチしないため、解析から除外されるリードペア。
方向曖昧断片	順鎖および逆鎖の両方向の転写産物にマッチするリードペア。
イントロン断片	遺伝子と重複するが、エクソンと重複しないリードペア。
遺伝子間断片	どの遺伝子にも重複しないリードペア。
未知転写産物断片	遺伝子のエクソンと重複するが、どの転写産物(ミスマッチしたスプライス部位)にもマッチしないリードペア。

RNAバリアントコール

DRAGEN RNAバリアントコールは、DRAGEN体細胞スモールバリアントコーラーを使用して、SNVおよびIndelをコールします。DRAGENは体細胞バリアントコールを使用して、発現差によって生じるRNA-Seqデータの非生殖細胞バリアントアリルを評価します。バリアントコールを実施するために、DRAGENは実際のバリアントエビデンスとさまざまなノイズモデルのエビデンスを比較検討する確度モデルを使用します。バリアントのクオリティスコアが特定の閾値を超える場合、そのバリアントは、PASSラベルの付いた出力VCFにレポートされます。また、DRAGENはweak_evidenceおよびbase_qualityなどのフィルターを適用します。これは、そのバリアントがパスコールと見なされるために必要な閾値に達していない場合を示すことがあります。DRAGEN DNA体細胞バリアントコールに関する詳細は、104 ページの 「体細胞モード」 を参照してください。

RNAバリアントコールによる強制ジェノタイピング（ForceGT）も使用できます。関心のあるバリアントを含むVCFを入力でき、出力VCFはその入力からアノテーションを含むすべてのバリアントを含むことになります。ForceGTは複雑なバリアントまたは長い欠損（50 bp超）のあるバリアントを正確にコールすることができない場合があります。複雑なバリアントとは、REFアリルをALTアリルに置換するために複数の置換、挿入または欠失イベントを必要とするバリアントです。

入力オプション

入力としてFASTQ、BAM、CRAMファイルを使用できます。DRAGEN体細胞バリアントコールを有効にするには、腫瘍ファイルとして入力ファイルに必ずマークを付けてください。オプションとして、より正確なスプライスジャンクションマッピングを行うために、GTFアノテーションファイル提供することができます。

FASTQ入力ファイルには次のコマンドラインオプションを使用します。

```
--tumor-fastq1=<fastq1_file> \
--tumor-fastq2=<fastq2_file> \
--RGID=<read_group_id> \
--RGSM=<read_group_sample_name> \
```

FASTQ入力ファイルのリストには次のコマンドラインオプションを使用します。

```
--tumor-fastq-list=<fq_list_file> \
--tumor-fastq-list-sample-id=<sample_id>
```

BAM入力ファイルには次のコマンドラインオプションを使用します。

```
--tumor-bam-input=<bam_file> \
--enable-map-align=false \
--enable-sort=false \
--enable-duplicate-marking=false
```

RNAバリアントコールの実行

RNAバリアントコールを有効にするには、`--enable-rna`および`--enable-variant-caller`をtrueに設定します。ForceGTを有効にするには、`--vc-forcegt-vcf <forcegt_vcf_file>`を使用します。

RNAバリアントコールは、PASSバリアントおよび、フィルターまたはエビデンスが弱いためにパスしなかったバリアントを含むVCFファイルを出力します。フィルターおよび追加コマンドラインオプションの詳細については、104 ページの 「体細胞モード」 を参照してください。

以下はRNAバリアントコールコマンドラインの例を示します。

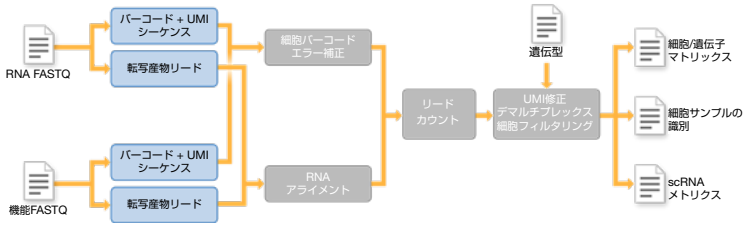
```
dragen \
--tumor-fastq1=<fastq1_file> \
--tumor-fastq2=<fastq2_file> \
--RGID=<read_group_id> \
```

```
--RGSM=<read_group_sample_name> \
--enable-duplicate-marking=true \
--dupmark-version=hash \
--enable-rna=true \
--enable-variant-caller=true \
--ref-dir=<ref_hashtable_dir> \
--output-directory=<output_dir> \
--output-file-prefix=<output_prefix> \
--annotation-file=<gtf_annotation_file> \
--vc-forcegt-vcf=<forcegt_vcf_file>
```

DRAGEN Single-Cell RNA Pipeline

DRAGEN Single-Cell RNA（scRNA）Pipelineはリードからcell-by-gene UMIカウントによる遺伝子発現マトリクスまでマルチプレックス化されたシングルセルRNA-Seqデータセットを処理できます。このパイプラインは、ある転写産物および細胞バーコードとUMIを含むその他の配列にマッチする断片中に1つリードがあるライブラリーデザインに適合します。このパイプラインには次の機能が含まれます：

- 転写産物リードに対してアノテーションされた遺伝子へのRNA-Seq（スプライス対応）アライメントおよびマッチング。
- 細胞バーコードとバーコードリードに対するUMIエラー補正。
- 遺伝子発現を測定するための細胞および遺伝子ごとのUMIカウンティング。
- 疎行列出力およびQCメトリクス。
- 細胞表面タンパク質を用いるなどの機能カウント。



アライメントおよび遺伝子アノテーションに関する機能性およびオプションはRNA-Seqパイプラインと同一です。詳細については、271 ページの 「DRAGEN RNA Pipeline」 を参照してください。遺伝子融合コールまたは転写産物レベルでの遺伝子発現定量などのその他のRNA-Seqモジュールはシングルセルでは

入力ファイル

アライメントリファレンス

DRAGEN scRNA Pipelineには標準的なDRAGEN RNAリファレンスゲノムまたはハッシュテーブルを使用します。例えば、`--ht-build-rna-hashtable=true`を使用して構築します。また、このパイプラインは*--annotation (-a)*オプションで提供される、GTF形式の遺伝子アノテーションファイルを必要とします。

リードインプット

DRAGEN scRNA Pipelineは、インプットとして各断片（リード）についての、転写産物配列とバーコード+UMI配列の両方を必要とします。転写産物配列は発現した遺伝子を特定するためにゲノムにアライメントされ、バーコード+UMI配列はユニークな細胞を同定するためのシングルセルバーコードと、ユニークな分子を定量するためのシングルセルUMIに分割されます。

FASTQから開始する場合、リード名にUMIを含めることも、別々のUMI FASTQファイルを提供することも可能です。

リード名中のUMI

リード名ラインの8番目のフィールドにバーコード+UMI配列を含むシングルエンドのFASTQファイルとしてトランスクリプトームリードを提供します。コロンを使って配列を分けます。以下の例は、リード2(`sample.R2.fastq.gz`)を転写産物リードとして使用しています。

```
@D00626:253:CAT5WANXX:4:1302:8433:85343:GAAACTCGTTCAGCGCTCATCTCGGC
ACAGGTCAGCTGGGAGCTTCTGCCCCCACTGCCTAGGGACCAACAGGGGCAGGAGGCAGTCACTGACCCCGAG
AAGTTTGCATCCTGCACAGCTAGAGATC
+
CCCCBFGGGGGGGGGGGGGGGGGGBDCGGGE>1BGDFGG0F@FBGGGBCDGGGGGGGGGGGGGGGGGGGGGG
DGGGGGGGGGGGGGFGGGGFGEGGGGGE
```

この例では、GAA配列は、バーコード+UMIリードであり、ACGC配列はトランスクリプトームリードです。

これらのFASTQファイルは、シングルセルバーコード+UMIリードを定義するためにUMI設定を用いてbclConvertおよびbcl2fastqによって生成されます。bclConvertを使用している場合、`OverrideCycles`設定を使用してバーコード/UMI情報を入力します。詳細については、『BCL Convert Software Guide』（文書番号：1000000094725）を参照してください。

bclConvertはシングルセルバーコード+UMI配列全体をUMIとして参照します。

bclConvertから生成されたFASTQファイルを使用するには、次のコマンドラインオプションを入力します：

```
dragen -1 <file name> --umi-source=qname
```

また、このオプションは--fastq-listインプットオプションおよびBAMファイルからのリードインプットとも適合します。

個別のUMI FASTQファイル

別のオプションでは、転写産物配列とバーコード+UMI配列を2つの別々のFASTQファイルとして提供します。1つのファイルはトランスクリプトームリードのみを含み、もう1つは同じ順序で一致するバーコードリードを含みます。このファイルの扱い方はリードペアの通常の扱い方と同様です。別々のUMIファイルを使用している場合、シーケンスシステムのランセットアップとbclConvertはUMIを全く認識せず、初期設定による通常のリード配列としてこのUMIを扱います。

別々のUMI FASTQファイルを使用するには、次のコマンドラインオプションを入力します：

```
dragen -1 <file name> --umi-fastq=<file name> --umi-source=fastq
```

複数のFASTQファイルにこのメソッドを使用するには、以下のようにしてください。

1. リード1としてfastq-listファイルにバーコード+UMI FASTQファイルを入力し、次にリード2としてbclConvertによって生成された初期設定fastq_list.csvに一致するトランスクリプトームリードのFASTQファイルを入力します。
2. 次のコマンドを入力します：

```
dragen --fastq-list fastq_list.csv --umi-source=read1
```

複数ライブラリーの使用

scRNAパイプラインはDRAGENランごとに単一の生物学的サンプルを処理できます。複数のシングルセルライブラリーを同時に処理するには、単一のサンプルを固有の細胞セットによって複数のシングルセルライブラリーにそれぞれ分割します。DRAGENは各ライブラリーからの細胞（バーコードとUMI）を別々に維持し、全体を統合した出力を提供します。リードグループは、RGLB属性を用いて各FASTQファイルに対するライブラリーを指定するために使用します。

シングルセルRNA設定

scRNAワークフローを使用するには、--enable-rna=true --enable-single-cell-rna=trueを入力します。本セクションは追加のscRNA設定に関する情報を掲載しています。

バーコード位置

初期設定では、scRNAワークフローはバーコード/UMI配列全体はシングルセルバーコード（複数のブロックに分割されている可能性があります）と単一のUMIで構成されていると想定しています。バーコードリードのシングルセルバーコードとシングルセルUMIの場所を同定するには、以下のコマンドラインを入力します：

```
--single-cell-barcode-position <blockPos>[+<blockPos2>+<blockPos3>...] --single-cell-umi-position <blockPos>
```

blockPosは1番目と最後のブロックに含める塩基のオフセットを説明しており、<startPos>_<endPos>として書式設定されています。16 bpの細胞バーコードの後に10 bpのUMIが続くライブラリーの例では、--single-cell-barcode-position 0_15 --single-cell-umi-position 16_25と入力します。固定したリンカー配列で区切られた3つのブロックの9 bpの細胞バーコードと8 bpのUMIを含むライブラリーでは、--single-cell-barcode-position=0_8+21_29+43_51 --single-cell-umi-position=52_59と入力します。

機能リードはある機能（例、細胞表面タンパク質または抗体）に特異的な配列タグのあるリードのことです。リード2に位置する機能特異的なUMIのある機能リードを含むscRNAランを実行している場合、`--single-cell-feature-barcode-r2umi=0_11`を使用して各機能リードの開始時点で12 bpの機能UMIを指定します。

バーコードの指定

次のコマンドの使用を含めるために、細胞バーコード配列のリストを提供することができます：

```
--single-cell-barcode-sequence-whitelist </path/to/barcodeAllowlist.txt>
```

ファイルにはラインごとに可能性のある細胞バーコード配列を1つ含める必要があります。gzip（*.txt.gz）でファイルを圧縮できます。細胞バーコードエラー補正中に、ファイルに指定した配列にマッチしない観察されたバーコードはエラーと見なされます。可能であれば、バーコードは同様の許可された配列に補正されます。詳細については、287 ページの 「バーコードエラー補正」 を参照してください。バーコードが補正されない場合、フィルタリングで除外されます。

細胞フィルタリング

DRAGENは細胞バーコードごとのユニークなUMI数の閾値を使用して、バックグラウンドノイズからどのバーコードがオリジナルサンプル中のシングルセルに一致する可能性があるかを特定します。閾値はバーコードごとのUMIの分布とサンプル中の実際の細胞の予想される数に基づいて決定されます。細胞フィルタリングの実施方法の詳細については、286 ページの 「細胞フィルタリング」 を参照してください。

- *single-cell-number-cells*：**（オプション）**予想される細胞数を設定します。初期設定は3000です。予想される細胞数が初期設定からかなり離れており、DRAGENが正しい細胞フィルタリング閾値を自動でコールしない場合にのみ調節します。
- *single-cell-threshold*：UMIカウント閾値を決定するためのメソッドを指定します。使用可能な値は`fixed`、`ratio`または`inflection`です。
  - `ratio`を使用している場合、DRAGENは`max(T_e, T_m)`として予想される細胞数を推測します。`T_m`は、最も豊富な細胞バーコードに見られるUMIの断片に基づく閾値です。`T_e`は、最も少ないと予想される細胞の断片に基づく閾値です。
  - `inflection`を使用している場合、DRAGENは最もUMIが豊富な細胞の累積UMIカウント数の変曲点解析に基づくアルゴリズムを使用して閾値を決定します。
  - `fixed`を使用している場合、UMIカウント閾値が設定されるため、動的に推測されるのではなく、予想された細胞数がパスします。パスした細胞の正確な数はシングルセル数よりもわずかに多くなることがあります。これはUMIカウントと異なる細胞バーコードとの関連によるものです。

  前もって特定の細胞数を設定するには、次のコマンドを使用します：

```
--single-cell-threshold=fixed --single-cell-number-cells=X
```

このコマンドは、最上位の X 細胞と、同数のユニークな UMI を含む追加の細胞をパスさせるために、UMI 閾値を強制的に設定します。

その他のオプション

以下は、Single-Cell RNA Pipeline設定を構成するために使用できるその他のオプションです。

オプション	説明
rna-libary-type	ゲノムに関連する転写産物リードの方向を設定します。フォワードには`SF`、リバースには`SR`、unstrandedには`U`を入力します。初期設定はSFです。
single-cell-count-introns	遺伝子発現推定にイントロンリードを含めます。初期設定はfalseです。
qc-enable-depth-metrics	ランタイムを早めるためには`false`に設定して、詳細なメトリクスを無効にします。初期設定はtrueです。
bypass-anchor-mapping	性能を向上させるためには`true`に設定し、RNAアンカー（two-pass）マッピングを無効にします。初期設定はfalseです。

コマンドラインの例

以下は、DRAGEN Single Cell RNA Pipelineを実行するためのコマンドラインの例です。

```
dragen --enable-rna=true --enable-single-cell-rna=true --umi-source=fastq
--single-cell-barcode 0_15 --single-cell-umi 16_25 -r reference_
genomes/Mus_musculus/mm10/DRAGEN/8 -a reference_genomes/Mus_
musculus/mm10/gtf/gencode.vM23.annotation.gtf.gz -1 lib1_S7_L001_R2_
001.fastq.gz --umi-fastq lib1_S7_L001_R1_001.fastq.gz --RGID=1 --
RGSM=sample1 --output-dir=/staging/out --output-file-prefix=sample1
```

バーコードエラー補正

インプットリードからの細胞バーコード配列は、それぞれが認められた頻度および予想される細胞バーコード配列のオプションの許可リストに基づいてエラー補正されます。細胞バーコード配列は、多くても1つのミスマッチがある場合、別の細胞バーコードの隣接する配列と判断されます。細胞バーコード配列は次の状況で近接する細胞バーコード配列に対して補正されます。補正される際に、近接する細胞バーコードではなく、細胞バーコードのあるリードすべてが割り当てられます。シーケンスエラー補正スキームは、（Smith, Heger and Sudbery, 2020）[1]に述べられたディレクショナルアルゴリズムに類似しています。

- 近接している細胞バーコードはすべてのインプットリード内で2倍以上頻度が高くなります。
- 近接する細胞バーコードは細胞バーコード許可リストに載っていますが、オリジナルの細胞バーコードは載っていません。

シーケンスエラーに基づくUMIの過剰カウントを避けるために、UMIエラー補正は同一遺伝子にマッピングする同一細胞バーコードのあるすべてのリードに対して実施されます。別のUMIエラーと考えられるUMI配列はカウントされません。

[1]Smith, T., Heger, A. and Sudbery, I., 2020. UMI-Tools: Modeling Sequencing Errors In Unique Molecular Identifiers To Improve Quantification Accuracy. [PDF] Cold Spring Harbor Laboratory Press. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5340976/> [Accessed 15 October 2020].

遺伝型サンプルのデマルチプレックス

DRAGENはいくつかの方法を実施して、1回のライブラリー調製やマイクロフルイディクスでプールされた細胞など、異なる個体の細胞が混在するデータセットをデマルチプレックスします。これらの方法のうちの2つは、遺伝型ベースと遺伝型フリーのデマルチプレックスがあります。遺伝型のデマルチプレックス法では、DRAGENは各細胞のリードに観察されたアリルに基づいて、細胞にサンプルIDを割り当てることができます。DRAGENはSNVに対してのみ実行します。DRAGENは異なる個体の複数の細胞を含むドロップレットなど、あらゆるダブレットにフラグを付けます。

遺伝型ベースのサンプルデマルチプレックスを使用するには、データセットに各サンプルに対する遺伝型を含んだVCFファイルを提供する必要があります。遺伝型フリーのサンプルデマルチプレックスを使用するには、できれば同一の遺伝的バックグラウンドのある集団からの外部サンプルセットを含むVCFファイルを提供する必要があります。`GT`フィールドはサンプル遺伝型を表しています。

細胞ハッシングのデマルチプレックス法に関する情報は、290 ページの 「細胞ハッシング」 を参照してください。

コマンドラインオプション

scRNAデマルチプレックスには、次のコマンドラインオプションを使用します。

オプション	説明
`--single-cell-demux-sample-vcf`	遺伝型ベースのサンプルデマルチプレックスを使用している場合、サンプル遺伝型を含むVCFファイルを指定します。
`--single-cell-demux-reference-vcf`	遺伝型フリーのサンプルデマルチプレックスを使用している場合、使用しているサンプルと同様の遺伝的バックグラウンドのある集団の遺伝型を含むVFCファイルを指定します。
`single-cell-demux-detect-doublets`	遺伝型ベースのサンプルデマルチプレックスでのダブレット検出を有効にします。初期設定値は`false`です。
`--single-cell-demux-number-sample`	使用しているサンプル数です。このオプションは`single-cell-demux-reference-vcf`オプションで指定した外部VCFリファレンス使用する時にのみ適用します。

以下は、遺伝型ベースのデマルチプレックスを用いてDRAGEN Single Cell RNA Pipelineを実行するためのコマンドラインの例です。

```
dragen --enable-rna=true --enable-single-cell-rna=true --umi-source=fastq
--single-cell-barcode 0_15 --single-cell-umi 16_25 -r reference_
genomes/Mus_musculus/mm10/DRAGEN/8 -a reference_genomes/Mus_
musculus/mm10/gtf/gencode.vM23.annotation.gtf.gz -1 lib1_S7_L001_R2_
001.fastq.gz --umi-fastq lib1_S7_L001_R1_001.fastq.gz --RGID=1 --
RGSM=sample1 --output-dir=/staging/out --output-file-prefix=sample1 --
single-cell-demux-detect-doublet=true --single-cell-demux-samplevcf=
sample.vcf
```

以下は、遺伝型フリーのデマルチプレックスを用いてDRAGEN Single Cell RNA Pipelineを実行するためのコマンドラインの例です。

```
dragen --enable-rna=true --enable-single-cell-rna=true --umi-source=fastq
--single-cell-barcode 0_15 --single-cell-umi 16_25 -r reference_
genomes/Mus_musculus/mm10/DRAGEN/8 -a reference_genomes/Mus_
musculus/mm10/gtf/gencode.vM23.annotation.gtf.gz -1 lib1_S7_L001_R2_
001.fastq.gz --umi-fastq lib1_S7_L001_R1_001.fastq.gz --RGID=1 --
RGSM=sample1 --output-dir=/staging/out --output-file-prefix=sample1 --
single-cell-demux-detect-doublet=true --single-cell-demux-referencevcf=
sample.vcf --single-cell-demux-number-samples=4
```

出力

以下の3ファイルでは、遺伝型ベースのscRNAサンプルデマルチプレックスの出力に関連した情報を得ることができます。

`<prefix>.scRNA.barcodeSummary.tsv`ファイルは細胞バーコードなど、細胞ごとのメトリクスを含みます。次の列は細胞ごとのデマルチプレックスに関する情報を含みます。`<prefix>.scRNA.barcodeSummary.tsv`メトリクスの詳細については、292 ページの 「シングルセルRNA出力」 を参照してください。

列	説明
SampleIdentity	`SampleIdentity`列は以下の値を含めることができます: • `sampleX`：特定の細胞（バーコード）はあるサンプルに一意に割り当てられます。 • `AMB(sampleX,sampleY)`：アルゴリズムはバーコードを割り当てるサンプルを特定できません。 • `MIX(mixing_coef*sampleX+(100-mixing_coef)*sampleY)`：細胞バーコードはダブレットとして分類されます。例えば：`MIX(50*sampleX+50*sampleY)`。

列	説明
IdentityQscore	IdentityQscore列はサンプルIDコールの信頼度を推測するために使用された値を含みます。DRAGENは`singlet`、`ambiguous`、または`doublet`として細胞のダブレット状況を特定した後、アイデンティティQスコアは-10 * log10として定義されます（2番目に可能性が高いアイデンティティとダブレット状況を仮定して、割り当てられたアイデンティティが正しい確率）。 より高い値のアイデンティティQスコアはより信頼度の高いサンプルIDコールと一致します。

`<prefix>.scRNA.demux.tsv`ファイルは各細胞のサンプルアイデンティティを推測するために使用されたサンプルデマルチプレックス統計量を含みます。

列	説明
Barcode	細胞と関連する細胞バーコード。
DemuxSNPCount	細胞バーコードのリードが交わるSNP数。
DemuxReadCount	少なくとも1つのSNPと交わる細胞バーコードのUMI数。
Pure Samples	VCFファイルからのサンプル。
BestMixtureIdentity	最高のログ尤度のある混合サンプル。`--single-cell-demux-detect-doublets=true`の場合にのみ使用可能です。
BestMixtureLogLikelihood	最高の混合サンプルのログ尤度。`--single-cell-demux-detect-doublets=true`の場合にのみ使用可能です。

`<prefix>.scRNA.metrics.demuxSamples.csv`ファイルは、`<prefix>.scRNA.metrics.csv`の全体のデータセットに対してレポートされるメトリクスと同様に、細胞ごとのメトリクスを含みます。

列	説明
Passing cells	パスした細胞バーコード数。
Fraction genic reads in cells	パスした細胞に割り当てられたカウント済みリード。
Median reads per cell	フィルターをパスした細胞あたりのカウント済みリードの合計。
Median UMIs per cell	フィルターをパスした細胞あたりのカウント済みUMIの合計。
Median genes per cell	フィルターをパスした細胞あたり1つ以上のUMIを含む遺伝子。

細胞ハッシング

DRAGENはいくつかの方法を実施して、1回のライブラリー調製やマイクロフルイディクスでプールされた細胞など、異なる個体の細胞が混在するデータセットをデマルチプレックスします。これらのメソッドの1つは、サンプルオリゴタグベースのメソッドであり、細胞ハッシングとして参照されます。

細胞ハッシングを使用するには、細胞ハッシングのCVSまたはFASTAリファレンスファイルを提供する必要があります。CVS形式では、機能バーコードリファレンスファイルは次のヘッダーを使用します：`id,name,read,position,sequence,feature_type`。

- `id`：その機能の識別子。例えば、ADT_A1018。
- `name`：その機能の名前。例えば、ADT_Hu.HLA.DR.DP.DQ_A1018。
- `read`：リード1（R1）またはリード2（R2）。
- `position`：機能バーコードの開始位置および長さを含む、指定したリード上の位置。例えば、`0_15`という位置は、位置0から開始し、長さが15である機能バーコードを表します。
- `sequence`：機能バーコードのDNA配列。例えば、`CAGCCCGATTAAGGT`。
- `feature_type`：その機能の種類。例えば：抗体キャプチャー。

コマンドラインオプション

細胞ハッシングサンプルのデマルチプレックスを有効にするには、次のコマンドラインオプションを指定します。

- `--single-cell-cell-hashing-reference`：サンプル特異的オリゴタグを含むCSVまたはFASTA細胞ハッシングリファレンスファイルを指定します。
- `--single-cell-demux-detect-doublets`：細胞ハッシングサンプルのデマルチプレックス中にダブレット検出を有効にします。初期設定値は`false`です。
- `--single-cell-demux-sample-fastq`：出力サンプル特異的FASTQファイル。詳細については、292 ページの 「サンプル特異的FASTQ出力ファイル」 を参照してください。

出力

`<prefix>.scRNA.barcodeSummary.tsv`ファイルは細胞バーコードなど、細胞ごとのメトリクスを含みます。`<prefix>.scRNA.barcodeSummary.tsv`ファイルの次の列には、細胞ごとの細胞ハッシング情報を含みます。`<prefix>.scRNA.barcodeSummary.tsv`ファイルに関する詳細については、292 ページの 「シングルセルRNA出力」 を参照してください。

列	説明
`SampleIdentity`	`SampleIdentity`列は以下の値を含めることができます: • `sampleX`：特定の細胞（バーコード）はあるサンプルにユニークに割り当てられます。 • `AMB(sampleX,sampleY)`：アルゴリズムはバーコードを割り当てるサンプルを特定できません。 • `MIX(mixing_coef*sampleX+(100-mixing_coef)*sampleY)`：細胞バーコードはダブレットとして分類されます。例えば：`MIX (50*sampleX+50*sampleY)`。

<prefix>.scRNA.demux.tsvファイルは各細胞のサンプルアイデンティティを推測するために使用されたサンプルデマルチプレックス統計量を含みます。

列	説明
Barcode	細胞と関連する細胞バーコード。
Pure samples	各サンプルに対する細胞ハッシングのリードカウント。

サンプル特異的FASTQ出力ファイル

サンプルデマルチプレックスアルゴリズムのいずれかを有効にした場合、各細胞のサンプルアイデンティティが利用可能になった後にサンプル特異的FASTQファイルを出力できます。次のコマンドラインを使用します。

```
--single-cell-demux-sample-fastq
```

gzipが指定されている場合、サンプル特異的出力FASTQファイルはgzip形式に圧縮されます。fastqが指定されている場合、サンプル特異的出力FASTQファイルは圧縮されません。初期設定オプションはnoneであり、これはサンプル特異的FASTQファイルが出力されることはないことを示します。

機能カウント

機能カウントを有効にするには、次のコマンドラインオプションを指定します：

- --single-cell-feature-barcode-reference：機能バーコードを含むCSVまたはFASTA機能リファレンスファイルを指定します。
- --single-cell-feature-barcode-groups：機能リードが別のFASTQファイルとして指定されている場合、FASTQ機能ファイルに一致するリードグループのカンマ区切りのリストを指定します。

機能カウントの出力は発現マトリクス出力に追加されます。拡張した発現マトリクスは機能に一致する追加の行が含まれます。

シングルセルRNA出力

シングルセルRNA出力はプレフィックスとしてscRNAを用いており、DRAGENの標準的な出力場所にあります。

カウント

以下の3ファイルはMatrix Market（*.mtx）形式で細胞ごとの遺伝子発現レベルの情報を提供します：

オプション	説明
<prefix>.scRNA.matrix.mtx	疎行列形式で各細胞/遺伝子ペアに対する一意のUMIをカウントします。
<prefix>.scRNA.barcodes.tsv	マトリクスからの各細胞に対する細胞バーコード配列。

オプション	説明
<prefix>.scRNA.genes.tsv	マトリクス中の各遺伝子に対する遺伝子名とID。これはすべての細胞バーコードが含まれます。パスした細胞と一致するサブセットは、scRNA.barcodeSummary.tsvのFilter列の下にあります。

アライメント

転写産物リードのアライメントは座標別にソートされ、BAMファイルとして出力されます。各アライメントは、細胞バーコードを含むXBタグおよびUMIを含むRXタグによってアノテーションされています。このアライメントはエラーが補正されていない、オリジナル列を使用します。関連したバーコードリードがなかった断片、例えばインプットデータでトリミングされた断片では、XBやRXタグはありません。

全体的なメトリクス

<prefix>.scRNA.metrics.csvファイルはサンプルあたりのscRNAメトリクスを含みます。

バーコードリードメトリクス

メトリクス	説明
Invalid barcode read	基本的なチェックを失敗した全バーコード配列（細胞バーコード+UMI）。例えば、バーコードリードが欠失しているまたは短すぎる。
Error free cell-barcode	エラー補正中に変化しなかった細胞バーコード配列のあるリード。例えば、リードが許可リストに完全マッチした場合。
Eallow listected cell-barcode	有効な配列への補正が成功した細胞バーコード配列のあるリード。
Filtered cell-barcode	有効な配列に補正されていない可能性のある細胞バーコード配列のあるリード。例えば、その配列が許可リストにマッチせず多くても1つのミスマッチがある。

転写産物リードメトリクス

メトリクス	説明
Unique exon match	固有遺伝子にマッチする有効な細胞バーコードとUMIのあるリード。

メトリクス	説明
Unique intron match	エクソンにマッチしないが、1つの遺伝子のイントロンに完全にマッチするリード。例えば、`--single-cell-count-introns=true`コマンドを使用している場合。
Ambiguous match	複数の遺伝子にマッチするリード。
Wrong strand	ライブラリータイプの定義とは反対側のストランド上の遺伝子に重なるリード。
Mitochondrial reads	マッチしている遺伝子がある場合、例えばミトコンドリアにマップするリード。
No gene	どの遺伝子にもマッチしないリード。`--single-cell-count-introns=true`を使用しない限り、イントロンリードを含みます。
Filtered multimapper	ゲノム中で複数のポジションアライメントするために除外されたリード。

UMIカウントメトリクス

メトリクス	説明
Total counted reads	固有遺伝子にマッチする有効な細胞バーコードとUMIのあるリード。
Reads with error-corrected UMI	別の類似するUMI配列にマッチさせるためにUMIがエラー補正されているカウント済みのリード。
Reads with invalid UMI	無効なUMI配列によってカウントされなかったリード。例えば、純粋なホモポリマーリード、またはNを含むリード。
Sequencing saturation	重複するUMIのあるリードの断片。1 -（UMI/リード）。
Unique cell-barcodes	カウントされたリード中のみの固有の細胞バーコード配列の全体数。
Unique UMIs	カウントされた固有の細胞バーコードとUMIを合わせた全体数。

細胞メトリクス

メトリクス	説明
UMI threshold for passing cells	フィルターをパスするために細胞バーコードに必要なUMI数。
Passing cells	フィルターをパスした細胞バーコード数。

メトリクス	説明
Fraction genic reads in cells	フィルターをパスした細胞に割り当てられたカウント済みリード。
Fraction reads putative cells	フィルターをパスした細胞に割り当てられたカウント済み全リード。
Median reads per cells	フィルターをパスした細胞あたりのカウント済みリードの合計。
Median UMIs per cells	フィルターをパスした細胞あたりのカウント済みUMIの合計。
Median genes per cells	フィルターをパスした細胞あたり1つ以上のUMIを含む遺伝子。
Total genes detected	フィルターをパスした1つ以上の細胞中の少なくとも1つのUMIを含む遺伝子。

細胞ごとのメトリクス

`<prefix>.scRNA.barcodeSummary.tsv`ファイルはエラー補正後の細胞ごとの各固有細胞バーコードに対する統計量のサマリーを含みます。

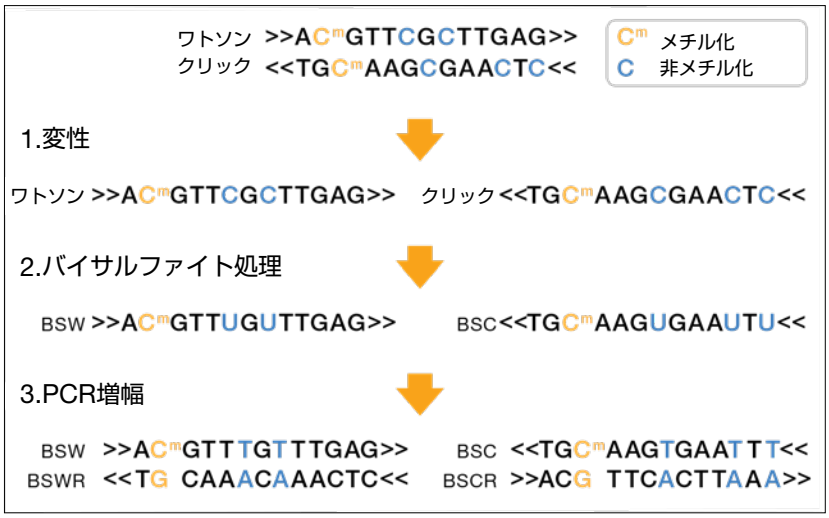
メトリクス	説明
ID	細胞バーコードに対する固有の数字ID。このIDはUMIカウントマトリクス(*.mtx)出力のラインに一致します。
Barcode	細胞バーコード配列。
TotalReads	細胞バーコード配列のあるリード合計。エラー補正されたリードを含みます。
GeneReads	遺伝子の方向をむいてカウントされたリード。
UMIs	カウントされたリード中のUMI総数。
Genes	検出された固有遺伝子。
MitochondrialReads	ミトコンドリアゲノムにマッピングされたリード。
Filter	使用できるフィルターは次のとおりです: • PASS：細胞バーコードはフィルターをパスします。 • LOW：UMIカウントは閾値未満です。

DRAGEN Methylation Pipeline

DNA中のシトシン塩基のエピジェネティックなメチル化は遺伝子発現に大規模な影響を及ぼし、バイサルファイトシーケンシングは1塩基レベルの解像度でエピジェネティックなメチル化のパターンを検出するために最もよく用いられる方法です。この技術は亜硫酸水素ナトリウムを用いてDNAを化学的に処理します。これにより非メチル化シトシン塩基はウラシルに変換されますが、メチル化されたシトシンは変換されません。その後のPCR増幅によってすべてのウラシルがチミンに変換されます。

バイサルファイトシーケンシングライブラリーはノンディレクショナルでもディレクショナルでも可能です。ノンディレクショナルライブラリーの場合、各二本鎖DNA断片から、以下の図に示すように、増幅後のシーケンス用に4種類の異なるストランドが生成されます。

図 14　ノンディレクショナルバイサルファイトシーケンシング



- バイサルファイトワトソン（BSW）、BSWの逆相補（BSWR）、
- バイサルファイトクリック（BSC）、BSCの逆相補（BSCR）

ディレクショナルライブラリーの場合、4種類のストランドが生成されますが、BSWとBSCストランドのみがシーケンスされるようにアダプターがDNA断片に付加されます（Listerのプロトコール）。一般的ではありませんが、BSWRおよびBSCRストランドがシーケンスに選択されます（例、PBAT）。

BSWおよびBSCストランド：

- A、G、T：変化なし
- メチル化CはCを維持
- 非メチル化CはTに変換

BSWRおよびBSCRストランド：

- 元のワトソン/クリックのA、G、T塩基に相補的な塩基は変化なし。
- 元のワトソン/クリックのメチル化Cに相補的なGはGを維持。
- 元のワトソン/クリックの非メチル化Cに相補的なGはAになる。

メチル化ハッシュテーブルの構築

DRAGEN Methylation Pipelineを実行するには、メチル化特異的なハッシュテーブルを構築する必要があります。マルチパスおよびシングルパスの両オプションにメチル化ハッシュテーブルが必要です。ハッシュテーブルの種類は、アライメント中に使用されるメソッドと一致している必要があります（`methylation-mapping-implementation: single-pass|multi-pass`）。マルチパスおよびシングルパスメソッドの詳細は、297 ページの 「マッピングメソッドのオプション」 を参照してください。

シングルパスメソッドを使用している場合、ハッシュテーブルは以下の要件を満たしていることを確認してください：

- `methylation-mapping-implementation: single-pass`を設定している。
- `methyl_converted`サブディレクトリに、統合したハッシュテーブルを生成するために`ht-methylated-combined=true`を設定している。
  元のFASTAファイルの各コンティグは組み合わせたリファレンスゲノムファイルに2回表示され、各変換タイプについて1回表示される。

マルチパスメソッドを使用している場合、ハッシュテーブルは以下の要件を満たしていることを確認してください：

- `methylation-mapping-implementation: multi-pass`を設定している。
- `CT_converted`および`GA_converted`サブディレクトリを生成するために`ht-methylated=true`を設定している。サブディレクトリは、CからT、GからAの変換に対するゲノムインデックスを含む。

`ht-methylated-combined=true`および`ht-methylated=true`を一緒に設定し、両方のメチル化マッピングメソッドに使用できるハッシュテーブルディレクトリを生成できます。

以下はシングルパスのハッシュテーブルに対するコマンドラインの例です。

```
dragen --build-hash-table true \
--output-directory=sample.output.directory \
--ht-reference=sample.input.fa \
--ht-num-threads 40 \
--ht-methylated=true \
--ht-methylated-combined=true \
--ht-seed-len 27
```

`--ht-seed-len 27`オプションは最適な結果を得るために重要です。`--ht-num-threads`オプションは、複数のスレッドを有効にして迅速に構築するために使用します。

マッピングメソッドのオプション

DRAGENはメチル化マッピング用に、マルチパスとシングルパスの2種類のメソッドに対応しています。

シングルパスメチル化マッピング

シングルパスメチル化マッピングは初期設定されています。シングルパスでは、DRAGENは1回のアライメントランのみ実行します。アライメント中、マッパーはリードに対して塩基とリファレンスの起こり得るすべての変換を検討し、変換が存在する場合、特定のメチル化ストランドに単一のベストなアライメントを出力します。検証したすべてのメチル化ストランド中に単一のベストなスコアリングアライメントがなかったリード（ペア）は、`MAPQ 0`としてBAM出力に表示されます。シングルパスのBAM出力には、XM、XRおよびXGメチル化タグのないマッピングされたリードが含まれる場合があります。メチル化タグのないリードからのメチル化データはレポートおよびメトリクスファイルに記録されません。

シングルパスメチル化マッピングを有効にするには、`--enable-methylation-calling`をtrueに設定する必要があります。

BAM出力にメチル化タグが含まれるようにし、レポートおよびメトリクスに記録されるようにするには、リードは以下の要件を満たす必要があります：

- リードおよびそのメイト（該当する場合）は、`--methylation-mapq-threshold`を用いて指定した値を上回るMAPQでマッピングされている。初期設定値は0である。
- リードは間違ったペアの一部ではない。

マルチパスメチル化マッピング

マルチパスメチル化マッピングでは、複数のアライメントランを連続して実行します。各ランからのアライメントは、ディスクに保存されます。すべてのアライメントランが完了した後、保存した結果は比較され、ベストなアライメントが選択されます。DRAGENは、検証したすべてのメチル化ストランド中に単一のベストなスコアリングアライメントがなかったリード（ペア）をBAM出力から除去します。BAM出力に表示されるすべてのリードにはメチル化BAMタグがあります。マッピング中のレポート生成を有効にしている場合、リードはメチル化レポートに含まれます。

マルチパスメチル化マッピングを有効にするには、`methylation-mapping-implementation=multi-pass`を設定します。

`--enable-methylation-calling`がtrueに設定されている場合、DRAGENは複数のアライメントを解析し、メチル化タグの付いた単一のBAMファイルを生成します。`--enable-methylation-calling`がfalseに設定されている場合、DRAGENはアライメントランごとに別々のBAMファイルを出力します。

`--methylation-match-bismark=true`が設定されている場合、CTOTまたはCTOBストランドにマッピングするペアエンドリードの位置はテンプレート内で反転されます。BAM出力にRead 1として表れるリードは2番目のFASTQファイルからのものになります。

DRAGENメチル化コール

異なるメチル化プロトコールは、インプットリードごとに2つまたは4つのアライメントを生成する必要があり、それを受けてベストなアライメントを選択し、シトシンのメチル化を決定するための解析を実施します。DRAGENは、単一のBAM出力ファイルと、メチル化コールおよびその他の下流のワークフローに使用できるBismarkと互換性のあるタグ（XR、XG、XM）を生成することでこの処理を自動化できます。

`--methylation-protocol`オプションがnone以外の有効な値に設定されている場合、DRAGENは必要とされるアライメントランのセットを自動的に生成します。各アライメントランには、リード上での適切な塩基変換、リファレンス上での塩基変換、およびリードがリファレンスに対してフォワード方向または逆相補（RC）方向にアライメントされなければならないかに対する制約が含まれます。

次のオプションは自動的に設定されます。

- --generate-md-tags true
- --Aligner.global 1
- --Aligner.no-unpaired 1
- --Aligner.aln-min-score 0
- --Aligner.min-score-coeff -0.2
- --Aligner.match-score 0
- --Aligner.mismatch-pen 4
- --Aligner.gap-open-pen 6
- --Aligner.gap-ext-pen 1
- --Aligner.supp-aligns 0
- --Aligner.sec-aligns 0
- --preserve-map-align-order true（マルチパスのみ）
- --seed-density 1（シングルパスのみ）

グローバルアライメント（リードの端から端まで）が生成されるため、DRAGENではライブラリー調製およびアダプター配列によって導入されたあらゆるアーティファクトのトリミングを推奨しています。

次の表はこれらのアライメントランを表します：

プロトコール	BAM	リファレンス	リード1	リード2	方向制約
ディレクショナル					
	1	C->T	C->T	G->A	フォワードのみ
	2	G->A	C->T	G->A	RCのみ
ノンディレクショナル、またはディレクショナル相補					
	1	C->T	C->T	G->A	フォワードのみ
	2	G->A	C->T	G->A	RCのみ
	3	C->T	G->A	C->T	RCのみ
	4	G->A	G->A	C->T	フォワードのみ

プロトコール	BAM	リファレンス	リード1	リード2	方向制約
PBAT					
	3	C->T	G->A	C->T	RCのみ
	4	G->A	G->A	C->T	フォワードのみ

ディレクショナルプロトコールでは、ライブラリーはBSWおよびBSCストランドのみがシーケンスされるように調製されます。その結果、アライメントの実行は、これらのストランド（上記ディレクショナルランの1と2）に最適な塩基変換と方向制約の2つの組み合わせによって実施されます。

ノンディレクショナルプロトコールでは、4ストランドそれぞれからのリードは同程度に調製されるため、アライメントの実行は、塩基変換と方向制約（上記ノンディレクショナルランの3と4）のさらに2つの組み合わせで実施される必要があります。

PBATプロトコールでは、BSWRおよびBSCRストランドのみがシーケンスされるようにライブラリーが調製されます。アライメントの実行は、塩基変換とこれらのストランド（ラン3と4）に最適な方向制約の組み合わせにより、2回のみ実施されます。

また、ディレクショナル相補プロトコールは、PBAT、または主にBSWRおよびBSCRストランドがシーケンスされる場合の同様のライブラリーに使用することもできます。このプロトコールでは、4回すべてのアライナーの実行がされますが、BSWおよびBSCストランド用のランからは相対的に良好なアライメントが少ないことが予想されるため、DRAGENはこれらのランに対してより高速の解析モードに自動で調整します。

以下はディレクショナルプロトコールに対するDRAGENコマンドラインの例です：

```
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
```

メチル化コール用のBismarkの使用

メチル化コールの推奨アプローチは、複数の必要なアライメントの実行を自動化し、XM、XR、XGタグを付加することです。マルチパスマッピングメソッドを使用している場合、`--enable-methylation-calling`をfalseに設定することでメチル化プロトコールに必要とされる制約と変換のそれぞれに対する別々のBAMファイルを生成できます。BAMファイルはメチル化コール用にBismarkにインプットとして使用できます。詳細は、イルミナテクニカルサポートにお問い合わせください。

このモードでは、1回のDRAGENの実行は`--output-directory`で指定されたディレクトリに複数のBAMファイルを生成します。各BAMファイルには、インプットリードと同じ順番のアライメントが含まれます。これらのランに対してソーティングまたは重複マーキングは有効にできません。アライメントはMDタグを含み、Bismark互換性のためにペアエンドリードの名前に`/1`または`/2`が付記されています。BAMファイルは次の命名規則を使用します：

- シングルエンドリード：output-directory/output-file-prefix.{CT,GA}read{CT,GA}reference.bam
- ペアエンドリード：output-directory/output-file-prefix.{CT,GA}read1{CT,GA}read2{CT,GA}reference.bam

ここで、output-directory（出力ディレクトリ）およびoutput-file-prefix（出力ファイルプレフィックス）は対応するオプションで指定され、CTおよびGAは上記の表に示された塩基置換に対応します。

Bismarkには、ディレクショナル相補モードはありませんが、ラン1とラン2がハイスコアのアライメントがほとんど生成されないことを予測して、Bismarkのノンディレクショナルモードを用いてそのようなサンプルを処理できます。

リードのソートおよび重複オプション

DRAGENはアライメントフェーズ中にメチル化されたリードに対するソーティングおよび重複のマーキング/除去に対応します。リードのソートおよび重複オプションを有効にするには、次の2つの異なるランを実施します。

1. 1番目のランに対するオプション設定は次のとおりです。
   - ソートされたアライメント出力（BAM形式）を生成するには、`--enable-sort`をtrueに設定します。
   - 重複リードを検出するには、`--enable-duplicate-marking`をtrueに設定します。
   - **（オプション）**重複リードを除去するには、`--remove-duplicates`をtrueに設定します。
   - `--methylation-generate-cytosine-report`および`--methylation-generate-mbias-report`をfalseに設定します。

   これらのオプションは DNA アライメントにおける場合と同じ動作をします。例えば、`--enable-duplicate-marking` が true に設定されている場合、`--enable-sort` は true です。
2. 2番目のランに対するオプション設定は次のとおりです。
   - `-b/--bam-input`に対して、これまでのランからソート/重複マーキング/重複除去のアライメント出力を使用します。
   - `--methylation-reports-only`をtrueに設定します。
   - `--enable-sort`をfalseに設定します。
   - シトシンレポートを生成するには、`--methylation-generate-cytosine-report`をtrueに設定します。
   - M-biasレポートを生成するには、`--methylation-generate-mbias-report`をtrueに設定します。

2番目のステップ中では、XM、XRおよびXGタグのあるリードからのメチル化塩基はレポートに記録されます。メチル化タグのないリードは無視されます。リードのソートおよび重複オプションがfalseに設定されている場合、アライメントファイル、シトシンレポートおよびM-biasレポートを生成するために必要なのはエンドツーエンドのランのみです。

初期設定では、DRAGENのメチル化解析はBismarkに従ってstrand-aware重複除去を実施します。ストランド対応の重複除去はマップされたリードをメチル化ストランドごとに1つずつの4グループに分割します。各グループ内でDRAGENは通常の重複除去を実施します。ペアリードの場合、そのペアの当該ストランドはそのペアの1番目のリードストランドとして定義されます。

以下の例はペアエンドリードに対するストランド対応の重複除去を示しています。例示したペアはすべて同じ位置にマッピングしていますが、XRタグとXGタグの異なる値で示されているように、各ペアの1番目のリード（BAMフラグ83および99）が異なるメチル化ストランドにマッピングされています。これらのペアは重複としてマークされません。

```
pair1 83 lambda 44001 60 150M = 43651 ... XR:Z:CT XG:Z:GA
pair1 163 lambda 43651 60 150M = 44001 ... XR:Z:GA XG:Z:GA
pair2 83 lambda 44001 60 150M = 43651 ... XR:Z:GA XG:Z:CT
pair2 163 lambda 43651 60 150M = 44001 ... XR:Z:CT XG:Z:CT
pair3 147 lambda 44001 60 150M = 43651 ... XR:Z:GA XG:Z:CT
pair3 99 lambda 43651 60 150M = 44001 ... XR:Z:CT XG:Z:CT
```

TAPSサポートの使用

TET-Assisted Pyridine Borane Sequencing（TAPS）は新しいアッセイであり、メチル化されたCを直接Tに変換します。一方、一般的なバイサルファイト変換はメチル化されていないCをTに変換します。このアプローチは、ゲノムの複雑性を保存し、有害な化学物質の使用を減少させ、インプットDNA量を減らすことができます。

TAPSから生成されるFASTQデータの解析を有効にするには、`--methylation-TAPS`をtrueに設定します。初期設定では、このオプションはfalseに設定されています。このオプションはアライメントステップ中にのみ実施され、メチル化シトシンおよび既存のBAMからのM-biasレポートを生成する際には必要ではありません。

メチル化関連BAMタグ

`--enable-methylation-calling`をtrueに設定した場合、DRAGENは構成した`--methylation-protocol`に対して生成されたアライメントを解析し、マッピングされたすべてのリードに対するメチル化関連タグを含む単一のBAM出力ファイルを生成します。Bismarkの場合、固有のベストアライメントの無いリードはBAM出力から除去されます。付加されるタグは次のとおりです。

タグ	簡単な説明	説明
XR:Z	リード変換	ベストアラインメントのために、リード時に実行された塩基変換：CTまたはGA。
XG:Z	リファレンス変換	ベストアライメントのために、リファレンスに対して実行された塩基置換：CTまたはGA。

タグ	簡単な説明	説明
XM:Z	メチル化コール	byte-per-baseのメチル化文字列。

XM:Z（メチル化コール）タグはリードの配列中の各塩基に対応するバイトを含みます。シトシンと関与しない各位置はピリオド（.）を含みます。シトシンと関与する各位置は文字を含みます。この文字はコンテキスト（CpG、CHG、CHHまたは不明）を示します。この場合はメチル化を指しています。メチル化された位置は大文字を使用し、非メチル化位置は小文字を使用します。シトシン位置で使用される文字は次のとおりです。

文字	メチル化されているか	コンテキスト
.	シトシンではない	シトシンではない
z	いいえ	CpG
Z	はい	CpG
X	いいえ	CHG
X	はい	CHG
h	いいえ	CHH
H	はい	CHH
u	いいえ	不明
U	はい	不明

メチル化シトシンおよびM-biasレポート

DRAGENを使用して、ゲノムワイドなシトシンメチル化レポートを生成できます。コマンドラインのオプション設定は、アライナーを介したFASTQを用いてランを実施している、またはメチル化タグを既に含むプレアライメントされたBAMを用いてランを実施しているかに応じて異なります。

- FASTQインプットには、`--methylation-generate-cytosine-report=true`を設定
- BAMインプットには、`--methylation-reports-only=true`を設定

`CX_report`のリファレンスからすべてのシトシンを維持しておくには、シトシンがインプットシーケンスに含まれていない場合であっても、`--methylation-keep-ref-cytosine true`を設定します。初期設定値はfalseです。このオプションをtrueに設定することで、ランタイムが長くなり、`CX_report`ファイルサイズが増加します。

シトシンレポートを圧縮するには、`--methylation-compress-cx-report = true`を設定します。初期設定値はfalseです。DRAGENは、`*.CX_report.txt`ではなく、圧縮された`*.CX_report.txt.gz`を出力します。

ゲノムの各Cの位置とストランドは、このレポートの初めの3つのフィールドに記載されます。ストランドのフィールドに – のある記録は、リファレンスのFASTAのGに対して使用されます。メチル化/非メチル化位置を含むメチル化Cおよび非メチル化Cの数は、4番目と5番目のフィールドに記載されます。リファレンスのCコンテキスト（CG、CHG、CHH）は6番目のフィールドに記載されます。トリヌクレオチド配列のコンテキストは最後のフィールドに記載されます（CCC、CGT、CGA等）。シトシンレポートはアライメントの間が1つ以上ある位置の記録のみを含みます。以下はシトシンレポート記録の例です：

```
chr2   24442367    +     18    0     CG    CGC
```

M-biasレポートを生成するには、`--methylation-generate-mbias-report`をtrueに設定します。このレポートには、シングルエンドデータに対する表が3つ、各Cコンテキストに対する表が1つ、ペアエンドデータに対する表が6つ含まれます。各表は一連の記録であり、リード塩基位置あたり1つの記録が含まれます。例えば、CHG表に対する1番目の記録は、1番目のリード塩基の位置に生じるメチル化C（フィールド2）と非メチル化C（フィールド3）の数を含み、1番目の塩基がゲノムのCHG部位にアライメントされるリードに限定します。また、表の各記録はメチル化C塩基の割合（フィールド4）とメチル化Cおよび非メチル化Cの数の総数（フィールド5）も含みます。

以下はリード塩基の位置10に対するM-bias記録の例です：

```
      10    7335    2356    75.69    9691
```

重複するペアエンドリードのあるデータセットの場合、シトシンレポートとM-biasレポートのどちらも1番目のリードに重複する2番目のリード中のCをレポートしません。また、1-based座標が両レポートの位置に対して使用されています。

マルチパスマッピングメソッドまたはレポートのみモードを使用している場合、`--methylation-match-bismark`オプションをtrueに設定することで、`bismark_ methylation_extractor`シトシンとBismarkバージョン0.19.0で生成したM-biasレポートを一致させることができます。BismarkおよびDRAGENのシトシンレポートの記録順序は異なる場合があります。DRAGENレポートはゲノム位置でソートされます。

出力メトリクス

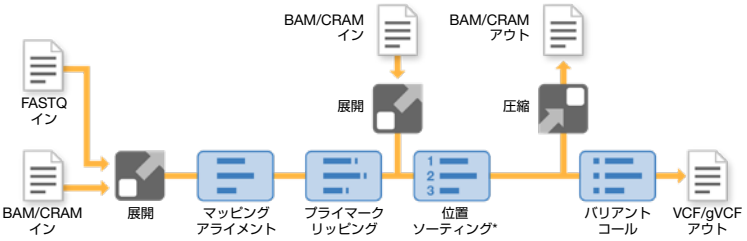
各メチル化ランの品質は次の2種類のメトリクスファイルに要約することができます。

- `*.mapping_metrics.csv`：合計リード数、アライメントリード数、重複除去リード数、塩基品質などのベンチマークを含む、アライメントフェーズに対して生成されるマッピング特異的なメトリクスを含みます。
- `*.methyl_metrics.csv`：解析したシトシン合計数、各シトシンコンテキスト中のメチル化率、ベストアライメントのストランドなどのベンチマークを含む、メチル化コールフェーズに対して生成されるメチル化特異的メトリクスを含みます。

DRAGEN Amplicon Pipeline

アンプリコンシーケンスは特定のゲノム領域の遺伝的変異解析を行うことができる高度にターゲットを絞ったアプローチです。PCR産物（アンプリコン）の超ディープシーケンスにより、バリアントの効率的な同定および特徴づけが可能です。この方法では、目的領域をターゲットにしてキャプチャーするために設計されたオリゴヌクレオチドプローブを使用した後、次世代シーケンサー（NGS）を使用します。

Amplicon PipelineはDNAデータのみに対応します。Amplicon Pipelineでは、ソフトクリッププライマーへのマッピングとアラインメントおよびアラインメントの書き換えの後に追加ステップを含めることにより、DRAGEN DNA Pipelineが使用します。ターゲットアンプリコンが見つかると、DRAGENは各アライメントをターゲットアンプリコンでタグ付けし、プライマー配列をソフトクリップします。DRAGENでは、出力BAM/CRAM記録に`XN:Z:<amplicon name>`タグを付加することによって、タグ付けを行います。ソフトクリップにより、プライマー配列がバリアントコールに関連しないようになります。Amplicon Pipelineでは、アッセイによるアンプリコンターゲットの性質により、各フラグメントは固有の開始位置と終了位置のみであるため、重複マーキングはオフになります。



*オプション

アンプリコンのBEDファイル

DRAGEN Amplicon Pipelineには、DRAGEN DNA Pipelineで必要とされるアンプリコンのBEDファイルとすべての入力ファイルが必要です。アンプリコンのBEDファイルの各行は、アンプリコンのターゲットについて説明しています。以下が必須フィールドです。

フィールド	説明
`chrom`	染色体の名称。
`chromStart`	プライマーを除いた、ターゲットの0-based inclusive開始位置。
`chromEnd`	プライマーを除いた、ターゲットの0-based exclusive終了位置。
`name`	アンプリコンのターゲットの名称。
`gene`	**(オプション)**遺伝子ID。

コマンドラインの設定

Amplicon Pipelineを使用するには、`--enable-dna-amplicon`を`true`に設定します。`--amplicon-target-bed`を使って、アンプリコンのBEDファイルのパスを指定します。

スモールバリアントコールのみが対象です。スモールバリアントコールを使用するには、`--enable-variant-calling`を`true`に設定し、`--vc-target-bed`を使って、ターゲットのスモールバリアントコールのBEDファイルをインプットします。ターゲットのBEDファイルをアンプリコンのBEDファイルに設定します。体細胞のスモールバリアントをコールする場合、`--vc-use-somatic-hotspots`を`false`に設定します。

初期設定では、アンプリコンのプライマーの最大長は25に設定されています。`--amplicon-primer-length`を使って、さまざまな値を指定することができます。アライメントがアンプリコンのターゲットに割り当てられるかどうかについて、パラメーターが影響を及ぼします。アライメントがアンプリコンのターゲットのプライマー領域内部から開始する場合、アライメントはアンプリコンに割り当てられます。適切なペアのアライメントの場合、アライメントとメイトは同じアンプリコンのターゲットからのものでなければなりません。

```
|-- primer --|-- amplicon target --|-- primer --|
  ---------- read ----------------->
     <---------- read -----------------
```

以下は、生殖細胞系列のスモールバリアントコールで、DRAGEN Amplicon Pipelineを実行するためのコマンドラインの例です。

```
dragen --enable-dna-amplicon true --enable-map-align=true --enable-
sort=true --enable-map-align-output=true -r reference_
genomes/Hsapiens/hg19_alt_aware/DRAGEN/8 --amplicon-target-
bed=CancerHotSpot-v2.dna_manifest.20180509.bed --enable-variant-
caller=true --vc-target-bed=CancerHotSpot-v2.dna_manifest.20180509.bed --
fastq-file1=read1.fastq.gz --fastq-file2=read2.fastq.gz --RGSM NA12878 --
RGID 1 --output-directory=/staging/out --output-file-prefix=NA12878
```

ツールとユーティリティ

BCL変換

DRAGENのBCL変換は、bcl2fastq2 v2.20アウトプットにマッチするFASTQファイルを出力するように設計されています。DRAGENデータ解析は、以下の機能をサポートします。

- 必要に応じて、ミスマッチ許容値を指定しながら、バーコードでサンプルをデマルチプレックス。
- マッチングを厳密に調整しながら、アダプターシーケンスのマスキングまたはトリミングをサポート。
- UMIシーケンスのタグ付けと必要に応じたトリミングをサポート。
- インデックスリード用FASTQファイルの出力をサポート。
- すべてのレーンを、同じFASTQ出力ファイルに結合。
- 大量のサンプル（100,000）をサポート。
- インデックスリードでUMIシーケンスをサポート。
- MinimumAdapterOverlap設定を使用して、アダプターシーケンスをトリミングした結果として生じるスキューを除去。
- デマルチプレックス、クオリティスコア、アダプタートリミング、マッピングされていないバーコード、インデックスホッピング検出のメトリクスを出力。
- 許可リスト、またはブロックリスト、もしくはその両方を使用して、正規表現で指定されたタイルのサブセットを変換。

BCLメトリクス

DRAGENデータ解析は、以下のメトリクスをCSV形式で、Reports/outputサブフォルダーに出力します。また、変換中に使用されたサンプルシートと`RunInfo.xml`ファイルもReports/outputサブフォルダーにコピーされます。

デマルチプレックス出力ファイル

以下のメトリクスは、`Demultiplex_Stats.csv`出力ファイルに含まれます。

列	説明
# One Mismatch Index Reads	バーコードを持つマッピング済みリードのうち、1塩基のミスマッチが存在するリードの数。
# Perfect Index Reads	サンプルシートに提供されたインデックスとマッチするバーコードを持つマッピング済みリードの数。

列	説明
Index	このサンプルのサンプルシートにあるindexの内容。デュアルインデックスの場合、値はindex2と結合されます。
# Reads	レーンのこのサンプルにマッピングされたパスフィルターリードの総数。
#Two Mismatch Index Reads	バーコードを持つマッピング済みリードのうち、厳密に2塩基のミスマッチが存在するリードの数。
% One Index Reads	バーコードを持つマッピング済みリードのうち、厳密に1塩基のミスマッチが存在するリードのパーセンテージ。
% Perfect Index Reads	サンプルシートに提供されたインデックスと厳密にマッチするバーコードを持つマッピング済みリードのパーセンテージ。
% Reads	レーンのこのサンプルにマッピングされたパスフィルターリードのパーセンテージ。
% Two Index Reads	バーコードを持つマッピング済みリードのうち、厳密に2塩基のミスマッチが存在するリードのパーセンテージ。
Lane	各メトリクスのレーン。
SampleID	このサンプルのサンプルシートにあるSample_IDの内容。

クオリティ出力ファイル

以下のメトリクスは、Quality_Metrics.csv出力ファイルに含まれます。

列	説明
% Q30	このリードのサンプルにマッピングされ、クオリティスコアが30以上の塩基のパーセンテージ。
index	このサンプルのサンプルシートにあるIndex 1 (i7)の内容。
index2	このサンプルのサンプルシートにあるIndex 2 (i5)の内容。
Lane	このメトリクスレーンが参照しているレーン番号。
Mean Quality Score (PF)	このリードのサンプルにマッピングされた塩基のクオリティスコアの平均値。
QualityScoreSum	このリードのサンプルにマッピングされた塩基のクオリティスコアの合計値。
ReadNumber	このメトリクスレーンが参照しているリードの番号。
Sample_ID	このサンプルのサンプルシートにあるSample_IDの内容。
Yield	このリードのサンプルにマッピングされた塩基の総数。
YieldQ30	このリードのサンプルにマッピングされ、クオリティスコアが30以上の塩基の総数。

アダプター出力ファイル

以下のメトリクスは、`Adapter_Metrics.csv`出力ファイルに含まれます。

列	説明
% Adapter Bases	アダプターとして、サンプルのリードからトリミングされた塩基のパーセンテージ。
AdapterBases	アダプターとして、サンプルのリードからトリミングされた塩基の総数。
index	このサンプルのサンプルシートにあるIndex 1 (i7)の内容。
index2	このサンプルのサンプルシートにあるIndex 2 (i5)の内容。
Lane	このメトリクスレーンが参照しているレーン番号。
ReadNumber	このメトリクスレーンが参照しているリードの番号。
Sample_ID	このサンプルのサンプルシートにあるSample_IDの内容。
SampleBases	サンプルのリードからトリミングされなかった塩基の総数。

インデックスホッピング出力ファイル

ユニークデュアルインデックスインプットの場合、`Index_Hopping_Counts.csv`ファイルは、ミスマッチ許容値を介するものを含め、指定されたindexとindex2の値の組み合わせごとにマッピングされるリードの数を提供します。メトリクスからは、発生したすべてのインデックスホッピングの動作がよくわかります。インデックスホッピングファイルには、サンプルシートに存在するindexとindex2の両方の値を持つサンプルがあります。`Index_Hopping_Counts.csv`出力ファイルには、以下の情報が含まれます。

列	説明
Lane	各メトリクスのレーン。
SampleID	インデックスのペアがサンプルに対応している場合、そのサンプルのサンプルシートにあるSample_IDの内容。
index	サンプルのサンプルシートにあるindexの内容。
index2	サンプルのサンプルシートにあるindexの内容。
# Reads	indexとindex2のペアにマッピングされたパスフィルターリードの総数。
% of Hopped Reads	indexとindex2のペアにマッピングされたパスフィルターリードのうち、ホッピングされたもののパーセンテージ。
% of All Reads	indexとindex2のペアにマッピングされたすべてのパスフィルターリードのパーセンテージ。

主な不明バーコード出力ファイル

`Top_Unknown_Barcodes.csv`ファイルには、フローセルインプットに頻繁に登場するバーコードシーケンスのうち、サンプルシートにリストアップされていないものが列挙されます。リストアップされていないシーケンスのうち、登場頻度の高い上位100個に加え、100番目のシーケンスと同等の頻度で出てくるその他のシーケンスも含まれます。`Top_Unknown_Barcodes.csv`出力ファイルには、以下の情報が含まれます。

列	説明
Lane	各メトリクスのレーン。
index	リストアップされていないシーケンスの1つめのindex値。
index2	リストアップされていないシーケンスの2つめのindex値。
# Reads	indexとindex2のペアにマッピングされたパスフィルターリードの総数。
% of Unknown Barcodes	indexとindex2のペアにマッピングされた不明パスフィルターリードのパーセンテージ。
% of All Reads	indexとindex2のペアにマッピングされたすべてのパスフィルターリードのパーセンテージ。

FASTQ出力ファイル

BCLファイルの変換バージョンとして、FASTQファイルはDRAGENの第1の出力です。BCLファイルと同様、FASTQファイルには、ベースコールと関連するQスコアが含まれます。しかし、サイクルごとのデータを含むBCLファイルとは異なり、FASTQファイルには、多くのデータ解析アプリケーションで必要とされる、リードごとのデータが含まれます。

DRAGENは、サンプル、リード、レーン1つにつき、FASTQファイルを1つ生成します。例えば、ペアエンドランのサンプル1つについて、Read 1用に1つ、Read 2用に1つ、合計2つのFASTQファイルを生成します。これらのサンプルFASTQファイルに加えて、DRAGENは、すべての不明サンプルを含むFASTQファイルをレーン1つにつき2つ生成します。Index Read 1とIndex Read 2のためにFASTQファイルが生成されることはありません。これは、このシーケンスが、各FASTQエントリーのヘッダーに含まれているからです。

`fastq_list.csv`出力ファイルは、FASTQファイルとともに、出力フォルダーに保管されます。DRAGENは、サンプルインデックス、レーン、FASTQ出力ファイル名を関連付けます。各行の列は、テストランのサンプルエントリーとともに表示されます。`fastq_list.csv`を使ったDRAGENの実行について、詳しくは「FASTQ CSVファイル形式」を参照してください。

列	説明
RGID	リードグループ

列	説明
RGSM	サンプルID
RGLB	ライブラリー
Lane	フローセルレーン
Read1File	有効なFASTQ入力ファイルへのフルパス
Read2File	有効なFASTQ入力ファイルへのフルパス。ペアエンドインプットでは必須です。ペアエンドインプットを使用しない場合は、空欄のままにしてください。

以下はfastq_list.csv出力ファイルの例です。

```
RGID,RGSM,RGLB,Lane,Read1File,Read2File
AACAACCA.ACTGCATA.1,1,UnknownLibrary,1,/home/user/dragen_bcl_out/1_S1_
L001_R1_001.fastq.gz,/home/user/dragen_bcl_out/1_S1_L001_R2_
001.fastq.gz
AATCCGTC.ACTGCATA.1,2,UnknownLibrary,1,/home/user/dragen_bcl_out/2_S2_
L001_R1_001.fastq.gz,/home/user/dragen_bcl_out/2_S2_L001_R2_
001.fastq.gz
CGAACTTA.GCGTAAGA.1,3,UnknownLibrary,1,/home/user/dragen_bcl_out/3_S3_
L001_R1_001.fastq.gz,/home/user/dragen_bcl_out/3_S3_L001_R2_
001.fastq.gz
GATAGACA.GCGTAAGA.1,4,UnknownLibrary,1,/home/user/dragen_bcl_out/4_S4_
L001_R1_001.fastq.gz,/home/user/dragen_bcl_out/4_S4_L001_R2_
001.fastq.gz
```

FASTQファイルのディレクトリ

このソフトウェアは、圧縮され、デマルチプレックスされたFASTQファイルを、コマンドライン`--output-directory`で定義されたディレクトリに書き込みます。

素性のわからないインデックスを持つリードは、`Undetermined_S0_`という名前のファイルに記録されます。サンプルシートの1レーンに複数のサンプルが含まれている場合、インデックスを指定する必要があります。サンプルシートの1レーンに複数のサンプルが含まれていない場合、ソフトウェアはバーコード不明エラーを表示し、データ解析を終了します。

ファイル名の形式は、サンプルシートで指定されたフィールドから構成されます。形式は`<Sample_ID>_S#_L00#_R#_001.fastq.gz`です。

このソフトウェアでは、インデックスされていないサンプルを1つ持つことができます。これは、1つのサンプルをシーケンスするのに同定は必要ないからです。しかし、複数のサンプルをシーケンスするには、マルチプレックスを行い、サンプルを同定して、解析できるようにする必要があります。

--no-lane-splittingオプションが有効である場合、レーンの表示はファイル名から削除されます。例えば、<Sample_ID>_S#_<R or I>#_001.fastq.gzのようになります。

ファイル名	説明
Sample_ID	そのサンプルのサンプルシートで指定された列エントリー。サンプルへデマルチプレックスされないパスフィルターリードには、Undeterminedが割り当てられます。
S#	サンプル数。この値は、サンプルシートにあるエントリーのレーン依存順に対応します。あるSample_IDが、異なるインデックスを持つレーンに複数回現れる場合は、同じS#が使用されます。サンプルへデマルチプレックスされないパスフィルターリードには、S0が割り当てられます。
L00#	そのサンプルのサンプルシートで指定された列レーン番号。
<R or I>	RunInfo.xmlの指定にしたがって、インデックスなし(R)またはインデックス付き(I)のリードタイプを表します。#は、RunInfo.xmlにこのリードタイプが現れる順番で、#には12が入ることがあります。

コマンドラインオプション

DRAGENのデータ解析は、以下のオプションを使って制御します。

ソフトウェアの動作

オプション	説明	初期設定
--bcl-conversion-only-true	DRAGEN実行可能ファイルでのFASTQファイルへのBCL変換に必要です。	該当なし
--bcl-input-directory	メインのコマンドラインオプションで、ランフォルダーディレクトリへのパスを表します。	該当なし
--bcl-only-lane	**(オプション)** 指定されたレーン番号のみ変換します。値は、RunInfo.xmlで指定されたレーン数以下でなければなりません。単一の整数値でなければなりません。	RunInfo.xmlで指定されたとおりすべてのレーン
--bcl-only-matched-reads-true	マッピングされていないリードがUndeterminedとマークされたFASTQファイルにアウトプットされないようにします。	false

オプション	説明	初期設定
--bcl-sampleproject-subdirectories	**（オプション）** trueの場合、サンプルシートに指定されたとおり、Sample_Projectサブディレクトリの作成を許可します。データセクションで、Sample_Project列を使用するには、使用するデータセクションのSample_Project列で、このオプションをtrueに設定する必要があります。	false
--bcl-use-hw-false	DRAGENデータ解析でBCL変換を同時に実行できるようにします。 BCL変換中、DRAGEN FPGAアクセラレーションを使用してはいけません。	該当なし
--exclude-tiles	タイルが正規表現と一致する場合、そのタイルが--tilesに含まれていても変換しません。	該当なし
--first-tile-only	**（オプション）** trueの場合、このオプションは、サンプルシートで指定された各レーンの最上位サーフェスにある先頭スワスの先頭タイルのみ処理します。 falseの場合、サンプルシートで指定されたとおり、各レーンにあるすべてのタイルを処理します。	false
--no-lane-splitting	レーン全体でFASTQファイルを統合します。 個々のサンプルは、リードごとに、同一のファイルへ提供されます。 • trueまたはfalseでなければなりません。 • Lane列が、サンプルシートから除外されている場合のみ、指定できます。	false
--no-lane-splitting-true	フローセルのレーンをすべて、同一のFASTQファイルへ連続的にアウトプットします。	false
--output-directory	必須のコマンドラインオプションで、デマルチプレックスされたFASTQアウトプットへのパスを示します。-f --forceが指定されていない限り、このディレクトリが存在していてはいけません。	該当なし
--run-info	RunInfo.xmlファイルへのパスをオーバーライドします。	--bcl-input-directory

オプション	説明	初期設定
`--sample-sheet`	**（オプション）**初期設定と異なる場合、サンプルシートの場所と名前を表すパスを示します。	<--bcl-inputdirectory>/SampleSheet.csv
`--strict-mode`	**（オプション）** `true`の場合、フィルター、LOCSIocs、BCLbcl、BCLbcレーンファイルのいずれかが所在不明、または破損している場合に、プログラムを中止します。 `false`の場合、フィルター、LOCSIocs、BCLbcl、BCIbciレーンファイルのいずれかが所在不明である場合に、プログラムを中止します。所在不明、または破損したファイルそれぞれについて、警告メッセージを表示します。	false
`--tiles`	正規表現と一致するタイルだけを変換します。	該当なし
`-f --force`	**（オプション）**`--output- directory`オプションで指定されたディレクトリの存在を許可します。	該当なし
`-h, --help`	ヘルプメッセージを表示して、アプリケーションを終了します。	該当なし
`-V, --version`	ソフトウェアバージョンを表示して、アプリケーションを終了します。	該当なし

ソフトウェアの性能

以下は性能を人為的にコントロールするために使用できるオプションです。このオプションは、共有マシンでDRAGENを実行しているときに、コアカウントを減らすためにのみ使用します。それ以外の場合は、必ず初期設定を使用してください。

CPUヘビーなスレッドの合計は、他のプロセスで使用するHWコア数未満でなければなりません。DRAGENで使用されるCPUヘビーなスレッドの総数は、以下の式で求められます：

$$(--bcl-num-parallel-tiles * --bcl-num-conversion-threads + --bcl-num-compression-threads + --bcl-num-decompression-threads)$$

オプション	説明	初期設定
`--bcl-num-parallel-tiles`	FASTQファイルに同時に変換されるタイルの数を指定します。指定できるのは1以上、利用可能なハードウェアスレッド数以下の値です。 32 CPUスレッドの場合は、`--bcl-num-parallel-tiles 2`を使用します。	動的に決定されます。CBCIインプットを使用している場合の値は4です。 集約BCLインプットを使用している場合の値は1です。
`--bcl-conversion-threads`	1タイルにつき、変換に使用するスレッドの数を指定します。指定できるのは1以上、利用可能なハードウェアスレッド数以下の値です。 32 CPUスレッドの場合は以下を使用します。 `--bcl-num-conversion-threads 4.`	動的に決定されます。サンプル数が10000を超える場合、値は1です。 サンプル数が10000未満の場合の値は、nproc / #parallel-tilesです。
`--bcl-num-compression-threads`	FASTQ出力ファイルの圧縮に使用するCPUスレッド数を指定します。指定できるのは1以上、利用可能なハードウェアスレッド数以下の値です。 32 CPUスレッドの場合は以下を使用します。 `--bcl-num-compression-threads 16.`	動的に決定されます。
`--bcl-num-decompression-threads`	入力ベースコールファイルの展開に使用するCPUスレッド数を指定します。指定できるのは1以上、利用可能なハードウェアスレッド数以下の値です。 32 CPUスレッドの場合は以下を使用します。 `--bcl-num-decompression-threads 8.`	動的に決定されます。
`--shared-thread-odirect-output`	実験的な共有スレッドファイル出力コードを使用します。これには`O_DIRECT`モードが必要です。trueまたはfalseでなければなりません。このファイル出力メソッドは、サンプル数が100000を超える場合に最適化されています。サンプル数がこれを下回る場合、またはアウトプット先にGPFSやLustreなどの分散ファイルシステムを使用している場合には、このオプションは推奨されません。	false

共有マシンで使用するコア数を減らすときには、CPUスレッドの数だけを調整します。

タイルフィルタリング

DRAGENデータ解析では、2種類のコマンドラインオプションを使って、変換するタイルを制御します。適切なコマンドラインは、タイルリストの発現によって異なります。

- `--tiles`：データ解析に含めるタイルを指定します。
- `--exclude-tiles`：データ解析から除外するタイルを指定します。

この機能は、bcl2fastq2の`tiles`、`ExcludeTiles`、`ExcludeTilesLaneX`の代わりです。`--tiles`と`--exclude-tiles`はどちらもタイル名（1つの正規表現）形式を使用します。例えば：

- `--tiles 1101`：すべてのレーンの先頭サーフェスと先頭スワスにある先頭タイルを含めます。
- `--tiles 11101`：NextSeq 500/550システムにのみ対応します。すべてのレーンの先頭サーフェスと先頭スワスにある先頭タイルを含めます。
- `s_`：レーンの指定に使用します。例えば：
  - `--tiles s_2`：レーン2の全タイルを変換します。
  - `-exclude-tiles s_2_1101`：レーン2の先頭サーフェスと先頭スワスにある先頭タイルを除外します。

範囲指定をする場合は、角カッコとハイフンを使って、指定します。例えば：

- `[1-2]101`：フローセルの両側の先頭タイルを選択します。
- `s_[1-2]_[1-2][0-9][0-9][0-9]5`：レーン1～2のサーフェス1～2と全スワスから5で終わる全タイルを選択します。

複数のタイル発現を指定するには、+を使います。

- `s_1_1102+s_[2-8]`：レーン1の先頭サーフェスと先頭スワスの2番目のタイルと、レーン2～8の全タイルを含めます。

`tiles`で使用する正規表現のあらゆる構成部分（+で区切られた部分）は、インプット`RunInfo`タイルリストにあるタイルエントリーの1つ以上と一致しなければなりません。`exclude-tiles`のあらゆる項は、同時に`tiles`オプションが使用されている場合はそれにより作成されたセット内のタイルエントリーの1つ以上と、または`RunInfo`タイルリストにあるタイルエントリーの1つ以上と一致しなければなりません。

サンプルシート

サンプルシート（`SampleSheet.csv`）には、サンプルに関する情報、対応するインデックス、DRAGENの動作を書きとったその他の情報が記録されます。サンプルシートのデフォルト位置はルートのアウトプットフォルダーです。別の場所やCSVファイルを指定するには、`--sample-sheet`コマンドを使用します。サンプルシートが初期設定の場所に存在せず、コマンドラインでも指定されていない場合、DRAGENはエラーになります。

BCL変換の動作をコントロールするコマンドラインオプションに加えて、サンプルシート構成ファイルの[Settings]セクションを使用して、サンプルの処理方法の指定することができます。BCL変換のサンプルシート設定は以下のとおりです。

サンプルシートのバージョン

DRAGENはサンプルシートv1とv2の両方をサポートしています。v1とv2でサポートされているオプションの違いを以下の表にまとめます。

サンプルシートv1	サンプルシートv2
[Settings]と[settings]の両方をサポートしています。どちらも必須ではありません	[BCLConvert_Settings]のみサポートしています。必須です。
認識されていない設定を使用すると、警告がトリガーされます。	認識されない設定を使用すると、エラーが発せられ、データ解析が中止されます。

Settingsセクション

DRAGENは、アダプタートリミング、サイクル、UMI、インデックスなどのオプションの指定に、サンプルシートの設定セクションを使用します。

表 10　アダプタートリミングの仕様

設定	初期設定	値	説明
`AdapterBehavior`	trim	trim、mask	DRAGENが、Read 1および/またはRead 2アダプター配列をマスキングするか、トリミングするかを定義します。AdapterRead1またはAdapterRead2が指定されていない場合、この設定を指定することはできません。 • mask：DRAGENは、同定されたRead 1および/またはRead 2配列をNでマスキングします。 • trim：DRAGENは、同定されたRead 1および/またはRead 2配列をトリミングします
`AdapterRead1`	該当なし	A、C、G、またはTを含むRead 1アダプターシーケンス。	マスキングまたはトリミングされるRead 1アダプター配列。 複数のアダプターをトリミングするには、各リードでマスキングまたはトリミングの評価が必要なアダプターを個別に示すため、配列をプラス記号(+)で区切ります。 使用可能な文字:A、T、C、G。

設定	初期設定	値	説明
AdapterRead2	該当なし	A、C、G、またはTを含むRead 2アダプターシーケンス。	マスキングまたはトリミングされるRead 2アダプター配列。 複数のアダプターをトリミングするには、各リードでマスキングまたはトリミングの評価が必要なアダプターを個別に示すため、配列をプラス記号(+)で区切ります。 使用可能な文字:A、T、C、G。
AdapterStringency	0.9	0.5～1.0の浮動小数点値	マスキングまたはトリミングをトリガーする最小マッチ率。この値は「マッチ数 / (マッチ数+ミスマッチ数)の式で計算されます。 許容値は0.5～1です。初期設定値0.9は、アダプターとの配列の同一性が90%以上のリードだけがトリミングされることを示します。
MinimumAdapterOverlap	1	1、2、または3	アダプターマッチが指定された塩基数以上でない限り、いかなる塩基もトリミングしません。MinimumAdapterOverlapを使用するには、AdapterRead1またはAdapterRead2を1つ以上指定する必要があります。 使用可能な文字:1、2、3。

表 11　サイクル、UMI、およびタイルの仕様

設定	初期設定	値	説明
BarcodeMismatchesIndex1	1	0、1、または2	index1で許容されるミスマッチ数。許容値は0、1、または2です。
BarcodeMismatchesIndex2	1	0、1、または2	index2で許容されるミスマッチ数。許容値は0、1、または2です。

設定	初期設定	値	説明
`CreateFastqForIndexReads`	0	0または1	インデックスリード用にFASTQファイルを出力するかどうかを指定します。インデックスリードがUMIと定義されている場合、DRAGENは、このUMI用にFASTQファイルを出力します（同時にTrimUMIも0に設定されている場合）。サンプルシートで、インデックスリードを1つ以上指定する必要があります。 • 0：FASTQファイルはインデックスリードのアウトプットではありません。 • 1：FASTQファイルは、FASTQリードのアウトプットです。
`MaskShortReads`	22	0～MinimumTrimmed ReadLength	アダプタートリミング後、A、T、C、Gの値を含む最小リード長。指定された塩基数に満たないリードはマスキングされます。値が21以下の場合、初期設定値がMinimumTrimmed ReadLengthになります。
`MinimumTrimmedReadLength`	35	0～インデックスなしリード長の最小値	アダプタートリミング後の最小リード長。DRAGENは、このパラメーターの値まで、リードからアダプターシーケンスをトリミングします。指定された値に満たない塩基はNでマスキングされます。

設定	初期設定	値	説明
NoLaneSplitting	false	trueまたはfalse	レーン全体でFASTQファイルを統合します。個々のサンプルは、リードごとに、必ず同一のファイルへ提供されます。 • trueまたはfalseでなければなりません。 • Lane列が、サンプルシートから除外されている場合のみ、指定できます。 • 指定されている場合、--no-lane-splittingコマンドラインパラメーターと同等です。 • trueの場合、フローセルのレーンをすべて、同一のFASTQファイルへ連続的にアウトプットします。

設定	初期設定	値	説明
OverrideCycles	Y:シーケンスリードを指定します I:インデックスリードを指定します U:リードからトリミングされるUMIの長さを指定します。	RunInfo.xmlで指定されたとおりにリードを使用します。	データの処理に使用するシーケンスサイクルとインデックスサイクルを指定します。 以下の形式を使用する必要があります: • 文字列内のセミコロンで区切られたフィールドの数は、RunInfo.xmlで指定されたシーケンスリードおよびインデックスリードと同じでなければなりません。 • インデックスリードはIで指定します。 • シーケンスリードはYで、UMIサイクルはUで指定します。 • トリミングされたリードはNで表します。 • 各リードに対して指定されたサイクル数は、RunInfo.xmlでそのリードに対して指定されたサイクル数の合計にならなければなりません。 • リード1つにつき、指定できるYまたはIシーケンスは1つだけです。 例: Y151;I8;I8;Y151

設定	初期設定	値	説明
TrimUMI	1	0または1	UMIサイクルをFASTQファイルから除外するかどうかを指定します。この設定を行うときには、サンプルシートに1つ以上のUMIを指定します。 • 0：UMIサイクルをFASTQファイルにアウトプットします。 • 1：UMIサイクルをFASTQファイルにアウトプットしません。

--no-lane-splitting true、または対応するサンプルシート設定NoLaneSplitting, trueを使用する場合、DRAGEN FASTQファイル命名規則とFASTQの内容は、同じ機能のbcl2fastq2と一致します。

レーン分割なし

--no-lane-splitting true、または対応するサンプルシート設定NoLaneSplitting- true,を使用する場合、DRAGEN FASTQファイル命名規則とFASTQの内容は、同じ機能のbcl2fastq2と一致します。

DRAGENがこのモードをサポートするのは、サンプルシートでLane列が指定されていない場合のみです。これにより、すべてのサンプルが、すべてのレーンに、リストされた順番で存在することを確認します。この順番は、レーン間に流路の境界を持たないフローセルで想定されます。

Dataセクション

Dataセクションは必須です。Dataセクションのヘッダーは、サンプルシートv1では[Data]または[data]、サンプルシートv2では[BCLConvert_Data]です。DRAGENは、Dataセクションの列を使用して、サンプルとインデックスアダプターをソートします。

列	説明
Lane	指定されている場合、DRAGENは、指定されたレーン番号のサンプルについてのみ、FASTQファイルを生成します。RunInfo.xmlで定義されたとおり、有効な整数を1つのみ指定できます。
Sample_ID	サンプルのIDです。
index	Index 1 (i7)インデックスアダプター配列。
index2	Index 2 (i5)インデックスアダプター配列。

列	説明
Sample_Project	使用できるのは英数字、ダッシュ、アンダースコアだけです。データ文字列の大文字小文字を変えて複製すること（例：sampleProjectとSampleProject）はできません。このようなデータ文字列を使用した場合、データ解析は失敗します。この列は、コマンドラインオプション--bcl-sampleproject-subdirectoriesを使用していない限り使用されません。コマンドラインオプションの詳細については、312 ページの「コマンドラインオプション」を参照してください。

使われなくなったサンプルシート設定

DRAGENは以下の設定をサポートしていません。必要に応じて、古い形式を、対応する新しい形式で置き換える必要があります。サンプルシートの[Settings]セクションは手作業で変更できますが、[Data]セクションはそのままにしておく必要があります。コマンドラインやサンプルシートで何らかの古い設定が使用されていた場合、DRAGENは停止し、エラーを返します。また、以前はコマンドラインで指定していた一部の設定を、サンプルシートで正しく指定できるようになりました。

表 12　アダプターの動作と仕様

動作	古い設定	新しい設定
Read 1とRead 2にアダプター配列を指定し、動作にトリミングを指定します。	（サンプルシート） Adapter, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA OR TrimAdapter, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA	（サンプルシート） AdapterRead1, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA AND AdapterRead2, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA

動作	古い設定	新しい設定
Read 1とRead 2に同じアダプター配列を指定し、動作にマスキングを指定します。	（サンプルシート） MaskAdapter, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA	（サンプルシート） AdapterRead1, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA AND AdapterRead2, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA AND AdapterBehavior, mask
Read 1とRead 2にアダプター配列を指定し、動作にマスキングを指定します。	（サンプルシート） MaskAdapter, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA OR MaskAdapterRead2, AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT	（サンプルシート） AdapterRead1, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA AND AdapterRead2, AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT AND AdapterBehavior, mask

動作	古い設定	新しい設定
Read 1とRead 2にアダプター配列を指定し、動作にトリミングを指定します。また、アダプター厳密度に0.5を指定します。	（サンプルシート） Adapter, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA OR TrimAdapter, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA （コマンドライン） --adapter-stringency 0.5	（サンプルシート） AdapterRead1, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA AND AdapterRead2, AGATCGGAAGAGCACACGTCTGAACTCCAGTCA（サンプルシート） AND AdapterStringency, 0.5

表 13　リードトリミング

動作	古い設定	新しい設定
151 x 8 x 8 x151ランで、Read 1の先頭から7個の塩基と最後から6個の塩基をトリミングします。	（サンプルシート） Read1StartFromCycle, 8 Read1EndWithCycle, 145	（サンプルシート） N7Y137N6;I8;I8;Y151

表 14　UMIの仕様

動作	古い設定	新しい設定
151 x 8x 8 x 151ランで、Read 1とRead2の先頭から8個のサイクルをUMIと指定し、後続塩基をトリミングします。	（サンプルシート） Read1UMIStartFromCycle, 1 Read1UMILength, 8 Read1StartFromCycle, 10 Read2UMIStartFromCycle, 1 Read2UMILength, 8 Read2StartFromCycle, 10	（サンプルシート） U8N1Y142;I8;I8;U8N1Y142

表 15　バーコードミスマッチ

動作	古いコマンドラインでの設定	新しいサンプルシートでの設定
i7インデックス配列とi5インデックス配列でそれぞれ1個のミスマッチを許容する	--barcode-mismatches 1 OR --barcode-mismatches 1,1	BarcodeMismatchesIndex1, 1 AND BarcodeMismatchesIndex2, 1
i7インデックス配列とi5インデックス配列でそれぞれ2個のミスマッチを許容する	--barcode-mismatches 2 OR --barcode-mismatches 2,2	BarcodeMismatchesIndex1, 2 AND BarcodeMismatchesIndex2, 2

表 16　トリミング後のリードのマスキング

動作	古いコマンドラインでの設定	新しいサンプルシートでの設定
アダプタートリミング後、9塩基対以下のリードすべてにNを追加して、トリミングされたリードがすべて、必ず10塩基対以上の長さになるようにします。	--minimium-trimmed-read- length, 10	MinimumTrimmedReadLength, 10
トリミング後、長さが4塩基対以下になったリードは、必ずNでマスキングします。	--mask-short-adapter-reads, 5	MaskShortReads, 5

Override Cycles

OverrideCyclesマスキング要素はセミコロンで区切ります。例えば：

```
OverrideCycles,U7N1Y143;I8;I8;U7N1Y143
```

DRAGENは、より多くのサードパーティアッセイをサポートするために、BCL変換中、インデックスリード中のUMI配列の存在や1リードあたり複数のUMI領域など、柔軟なUMI処理をサポートします。UMI配列は、FASTQリード配列からトリミングされ、通常どおり、各リードのシーケンス識別子に配置されます。

以下に、2x151リードを使用したOverrideCycles設定の例を示します：

設定	説明
OverrideCycles,U7N1Y143;I8;I8;U7N1Y143	UMIは、各ゲノムリードの先頭7 bpで構成され、これらは1 bpの無視される配列で繋がれています。また、イルミナのランダムではないUMIの形式で、以下の製品で使用されています： • TruSight Oncology 170 RUO • TruSight Oncology 500 RUO • IDT for Illumina - UMI Index Anchors
OverrideCycles,Y151;I8;U10;Y151	インデックスリード2は10 bpのUMIで、Agilent XT HSの形式です。
OverrideCycles,Y151;I8U9;I8;Y151	インデックスリード1は、インデックスを1つと9 bpのUMIを含み、UMIを使ったIDT Dual Index Adaptersの形式です。
OverrideCycles,U3N2Y146;I8;I8;U3N2Y146	UMIは、各ゲノムリードの先頭3 bpで構成され、これらは2 bpの無視される配列で繋がれています。また、SureSelect XT HS 2およびIDT xGen Duplex Seq Adapterで使用されているUMIの形式です。
OverrideCycles,Y151;I8;I8;U10N12Y127	UMIは、Read 2の先頭にあり、長さ12のリンカー配列で付加されています。

データ解析メソッド

DRAGEN Bio-It Platformは、以下のデータ解析を行います：

デマルチプレックス

DRAGENは、各レーンおよびリードのサンプル1つにつき1つのFASTQファイルを作成します。デマルチプレックスの動作は以下のとおりです。

- サンプルシートにマルチプレックス化されたサンプルが含まれている場合、DRAGENは以下の動作をします：
  - マッチするインデックスアダプター配列をつけずに、リードを`Undetermined_S0.fastq`にセットします。
  - 有効なインデックスアダプター配列をつけて、リードをサンプルFASTQファイルにセットします。
- サンプルシートに、インデックスのないサンプルが1つ含まれている場合、全リードがサンプルFASTQファイルにセットされます（Read 1とRead 2に1つずつ）。
- サンプルシートのDataセクションで定義されたサンプルにデマルチプレックスされないリードはすべて、レーンごとに`Undetermined_S0.fastq`にセットされます。
- DataセクションのLane列が使用されていない場合、すべてのレーンが変換されます。それ以外の場合は、入力されているレーンだけが変換されます。

UMIのトリミング

DRAGENは、ゲノムシーケンスまたはインデックスシーケンスから分子バーコード（UMI）配列をトリミングできます。UMIに対応するシーケンスリードのサイクルは、サンプルシートのSettingsセクションにあるOverrideCyclesパラメーターで指定します。OverrideCyclesパラメーターの設定については、317 ページの 「設定セクション」 を参照してください。

以下に、UMIと指定されたリードの動作の詳細をまとめます：

- 初期設定では、UMIは配列からトリミングされます。UMIを含めるには、サンプルシートでTrimUMI設定を使用します。
- UMI配列は、インデックスリードとゲノムリードで指定できます。リード1つにつき、複数のUMI配列を指定できます。
- 指定されたUMIサイクルは、すべてのクラスターに適用されます。レーンまたはサンプルに基づいて、UMIを適用するメカニズムはありません。
- UMI配列は、シーケンスリードおよびインデックスリードの先頭と末尾でのみ指定可能です。UMIをリードの途中に配置することはできません。

アダプターのトリミングとマスキング

DRAGENでは、リードデータでアダプター配列をマスキングまたはトリミングして、そのアダプター配列が下流の解析ステップに渡されないようにすることができます。

また、アダプターには以下のような処理能力もあります。

- DRAGENは、リード内のクラスター全体でリード長が一定になるように、同定されたアダプター配列をNでマスキングします。
- DRAGENは、同定されたアダプター配列を、リードからマスキングまたはトリミングします。トリミングされるため、各クラスターの長さは異なります。
- DRAGENは、インプットされるアダプター配列にはA、C、G、またはTのみ含まれていることを前提にしています。

システムの健全性のモニタリング

DRAGENシステムを立ち上げると、カードのハードウェア問題をモニタリングするデーモン（*dragen_mond*）が起動します。このデーモンは、DRAGENシステムをインストール、または更新したときにも起動します。その主な目的は、DRAGEN Bio-IT Processorの温度をモニタリングし、指定された閾値を上回ったときにDRAGENを停止することです。

モニタリングをマニュアルで起動、停止、または再起動するには、rootとして以下を実行します：

```
sudo service dragen_mond [stop|start|restart]
```

初期設定では、1分に1回、ハードウェア問題のポーリングが行われ、1時間に1回、温度がログに記録されます。

このサービスコマンドが実行されたときに*dragen_mond*の起動に使用されるコマンドラインオプションは、`/etc/sysconfig/dragen_mond`ファイルで指定します。初期設定オプションを変更するには、このファイルのDRAGEN_MOND_OPTSを編集します。例えば、ポーリング間隔を30秒、ログ記録時間を2時間に1回に設定するには、以下のように変更します：

```
DRAGEN_MOND_OPTS="-d -p 30 -l 7200"
```

*-d*は、モニタリングをデーモンとして実行するために必要なオプションです。

*dragen_mond*のコマンドラインオプションは以下のとおりです：

オプション	説明
-m --swmaxtemp <n>	ソフトウェアアラーム温度の最高値(摂氏)。初期設定は85です。
-i --swmintemp <n>	ソフトウェアアラーム温度の最低値(摂氏)。初期設定は75です。
-H --hwmaxtemp <n>	ハードウェアアラーム温度の最高値(摂氏)。初期設定は100です。
-p --polltime <n>	チップ状態レジスターをポーリングする間隔(秒)。初期設定は60です。
-l --logtime <n>	FPGA温度をn秒おきにログに記録します。初期設定はDefaultです。ポーリング間隔の倍数でなければなりません。
-d --daemon	デーモンとして分離し実行します。
-h --help	ヘルプを表示して、終了します。
-V --version	バージョンを表示して、終了します。

DRAGEN Bio-IT Processorの現在温度を表示するには、*dragen_info -t*コマンドを使用します。*dragen_mond*が実行されていない場合、このコマンドは稼動しません。

```
% dragen_info -t

FPGA Temperature: 42C (Max Temp: 49C, Min Temp: 39C)
```

ログへの記録

ハードウェアイベントはすべて、/var/log/messagesと/var/log/dragen_mond.logに記録されます。以下に示すのは、/var/log/messagesに記録された温度アラームの例です。

```
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
```

初期設定では、温度は1時間に一度、/var/log/dragen_mond.logに記録されます：

```
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)
```

DRAGENの実行中に温度アラームが検出された場合、DRAGENプロセスの端末ウィンドウに、以下のように表示されます：

```
**********************************************************
**  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.    **
**********************************************************
```

このメッセージが表示されたら、DRAGENソフトウェアの実行を停止してください。カードの過熱状態を軽減するために、以下を行います：

- カード上の通気が十分であることを確認します。より通気の良いスロットへカードを移動する、ファンを追加する、ファンの速度をあげるなどの対策を検討します。
- 筐体内でのカードのスペースを広げます。使用可能なPCIeスロットがあるならば、カードの両側に空のスロットがくるようにカードを移動します。

それでもシステムの温度アラームを解決できない場合には、イルミナのテクニカルサポートにお問い合わせください。

ハードウェアアラーム

アラームがトリガーされたときに、モニタリング機能により記録されるハードウェアイベントは以下のとおりです：

ID	説明	モニタリング機能の対応
0	ソフトウェアの過熱	DRAGEN Bio-IT Processorがソフトウェアの最低温度に下がるまで、使用を停止します
1	ハードウェアの過熱	致命的。DRAGENソフトウェアを中止します。システムのリブートが必須です
2	Board SPDの過熱	致命的ではないと記録されます
3	SODIMMの過熱	致命的ではないと記録されます
4	電源0	致命的。DRAGENソフトウェアを中止します。システムのリブートが必須です
5	電源1	致命的。DRAGENソフトウェアを中止します。システムのリブートが必須です
6	DRAGEN Bio-IT Processor電源	致命的ではないと記録されます
7	ファン0	致命的ではないと記録されます
8	ファン1	致命的ではないと記録されます
9	SE5338	致命的。DRAGENソフトウェアを中止します。システムのリブートが必須です
10–30	未定義（保留）	致命的。DRAGENソフトウェアを中止します。システムのリブートが必須です

致命的なアラームが発せられると、DRAGENホストソフトウェアは実行できなくなり、システムのリブートが必要になります。ソフトウェア過熱アラームがトリガーされた場合、モニタリング機能が実行中のDRAGENプロセスをすべて探し出し、中止します。モニタリング機能は、温度が下がって、最低閾値に到達し、ハードウェアがチップ状態アラームを解除するまで、新たなDRAGENプロセスをすべて中止し続けます。ソフトウェアの過熱アラームが解除されれば、DRAGENジョブの実行を再開できます。

これらのアラームのいずれかがシステムでトリガーされた場合は、イルミナのテクニカルサポートに連絡し、ログファイルの詳細をお知らせください。

Illumina Annotation Engine

Illumina Annotation Engine（IAE）、別名Nirvanaは、SNV、MNV、挿入、欠失、Indel、STR、SV、CNVなど、ゲノム変異について臨床レベルのアノテーションを提供します。インプットにはVCFを使用します。アウトプットは、すべてのアノテーションとVCFから抽出されたサンプル情報を構造化したJSON表現です。IAEは、複数のオルタネティブアリルと複数のサンプルに対処できます。

DRAGENでは、バリアントアノテーションソフトウェアをスタンドアロンで実装できます。

IAEを実行するには、以下の手順が必須です。

1. 外部データソース、遺伝子モデル、リファレンスゲノムをダウンロードします。
2. 結果として得られたJSONファイルにアノテーションを追加します。

初期設定では、IAEバイナリーは、`/opt/edico/share/nirvana`ディレクトリに配置されます。このディレクトリには、DownloaderとNirvana（Illumina Annotation Engine）の2つのファイルがあります。

制限

Illumina Annotation EngineとDownloaderは、以下のプラットフォームに対応しています：

- x64プロセッサーを使用するCentOS 6。
- x64プロセッサーを使用するCentOS 7とその他の最新Linuxディストリビューション。

データファイルのダウンロード

アノテーションデータファイルを格納するために、トップレベルディレクトリを作成します。作成したディレクトリには、以下の3つのサブディレクトリが含まれます：

- 遺伝子モデルを含むCache。
- dbSNPやgnomADのような外部データソースを含むSupplementaryAnnotation。
- リファレンスゲノムを含むReferences。

以下のコマンドラインオプションが使用されます：

オプション	値	例	説明
--ga	GRCh37、GRCh38、または両方	GRCh38	ゲノムアセンブリ
--out	Output directory	~/Data	トップレベルのアウトプットディレクトリ

以下のようにして、データファイルをダウンロードします。

1. データディレクトリを作成するには、以下のコマンドを入力します。
   この例では、ホームディレクトリの下にDataディレクトリが作成されます。
   ```
   mkdir ~/Data
   ```
2. ゲノムアセンブリで使用するファイルをダウンロードします。
   この例では、ゲノムアセンブリGRCh38をダウンロードします。
   ```
   /opt/edico/share/nirvana/Downloader --ga GRCh38 --out ~/Data
   ```

同じコマンドを使って、データソースとIllumina Annotation Engineサーバーを同期することができます。このとき、以下のアクションが行われます。

- 古いバージョンのデータソースなど、古くなったファイルをアウトプットディレクトリから削除。

- 最新のファイルをダウンロード。

作成されるアウトプットは以下のとおりです：

```
--------------------------------------------------------------------------
--
Downloader (c) 2020 Illumina, Inc.
Stromberg, Roy, Lajugie, Jiang, Li, and Kang 3.9.1-0-gc823805
--------------------------------------------------------------------------
--
- downloading manifest... 37 files.
- downloading file metadata:
- finished (00:00:00.8).
- downloading files (22.123 GB):
- downloading 1000_Genomes_Project_Phase_3_v3_plus_refMinor.rma.idx
(GRCh38)
- downloading MITOMAP_20200224.nsa.idx (GRCh38)
- downloading ClinVar_20200302.nsa.idx (GRCh38)
- downloading REVEL_20160603.nsa.idx (GRCh38)
- downloading phyloP_hg38.npd.idx (GRCh38)
- downloading ClinGen_Dosage_Sensitivity_Map_20200131.nsi (GRCh38)
- downloading MITOMAP_SV_20200224.nsi (GRCh38)
- downloading dbSNP_151_globalMinor.nsa.idx (GRCh38)
- downloading ClinGen_Dosage_Sensitivity_Map_20190507.nga (GRCh38)
- downloading PrimateAI_0.2.nsa.idx (GRCh38)
- downloading ClinGen_disease_validity_curations_20191202.nga (GRCh38)
- downloading 1000_Genomes_Project_Phase_3_v3_plus.nsa.idx (GRCh38)
- downloading SpliceAi_1.3.nsa.idx (GRCh38)
- downloading dbSNP_153.nsa.idx (GRCh38)
- downloading TOPMed_freeze_5.nsa.idx (GRCh38)
- downloading MITOMAP_20200224.nsa (GRCh38)
- downloading gnomAD_2.1.nsa.idx (GRCh38)
- downloading ClinGen_20160414.nsi (GRCh38)
- downloading gnomAD_gene_scores_2.1.nga (GRCh38)
- downloading 1000_Genomes_Project_(SV)_Phase_3_v5a.nsi (GRCh38)
- downloading MultiZ100Way_20171006.pcs (GRCh38)
- downloading 1000_Genomes_Project_Phase_3_v3_plus_refMinor.rma (GRCh38)
- downloading ClinVar_20200302.nsa (GRCh38)
- downloading OMIM_20200409.nga (GRCh38)
- downloading Both.transcripts.ndb (GRCh38)
- downloading REVEL_20160603.nsa (GRCh38)
```

```
- downloading PrimateAI_0.2.nsa (GRCh38)
- downloading dbSNP_151_globalMinor.nsa (GRCh38)
- downloading Both.sift.ndb (GRCh38)
- downloading Both.polyphen.ndb (GRCh38)
- downloading Homo_sapiens.GRCh38.Nirvana.dat
- downloading 1000_Genomes_Project_Phase_3_v3_plus.nsa (GRCh38)
- downloading phyloP_hg38.npd (GRCh38)
- downloading SpliceAi_1.3.nsa (GRCh38)
- downloading TOPMed_freeze_5.nsa (GRCh38)
- downloading dbSNP_153.nsa (GRCh38)
- downloading gnomAD_2.1.nsa (GRCh38)
- finished (00:04:10.1).
Description Status
------------------------------------------------------------------------
--
1000_Genomes_Project_(SV)_Phase_3_v5a.nsi (GRCh38) OK
1000_Genomes_Project_Phase_3_v3_plus.nsa (GRCh38) OK
1000_Genomes_Project_Phase_3_v3_plus.nsa.idx (GRCh38) OK
1000_Genomes_Project_Phase_3_v3_plus_refMinor.rma (GRCh38) OK
1000_Genomes_Project_Phase_3_v3_plus_refMinor.rma.idx (... OK
Both.polyphen.ndb (GRCh38) OK
Both.sift.ndb (GRCh38) OK
Both.transcripts.ndb (GRCh38) OK
ClinGen_20160414.nsi (GRCh38) OK
ClinGen_Dosage_Sensitivity_Map_20190507.nga (GRCh38) OK
ClinGen_Dosage_Sensitivity_Map_20200131.nsi (GRCh38) OK
ClinGen_disease_validity_curations_20191202.nga (GRCh38) OK
ClinVar_20200302.nsa (GRCh38) OK
ClinVar_20200302.nsa.idx (GRCh38) OK
Homo_sapiens.GRCh38.Nirvana.dat OK
MITOMAP_20200224.nsa (GRCh38) OK
MITOMAP_20200224.nsa.idx (GRCh38) OK
MITOMAP_SV_20200224.nsi (GRCh38) OK
MultiZ100Way_20171006.pcs (GRCh38) OK
OMIM_20200409.nga (GRCh38) OK
PrimateAI_0.2.nsa (GRCh38) OK
PrimateAI_0.2.nsa.idx (GRCh38) OK
REVEL_20160603.nsa (GRCh38) OK
REVEL_20160603.nsa.idx (GRCh38) OK
```

```
SpliceAi_1.3.nsa (GRCh38) OK
SpliceAi_1.3.nsa.idx (GRCh38) OK
TOPMed_freeze_5.nsa (GRCh38) OK
TOPMed_freeze_5.nsa.idx (GRCh38) OK
dbSNP_151_globalMinor.nsa (GRCh38) OK
dbSNP_151_globalMinor.nsa.idx (GRCh38) OK
dbSNP_153.nsa (GRCh38) OK
dbSNP_153.nsa.idx (GRCh38) OK
gnomAD_2.1.nsa (GRCh38) OK
gnomAD_2.1.nsa.idx (GRCh38) OK
gnomAD_gene_scores_2.1.nga (GRCh38) OK
phyloP_hg38.npd (GRCh38) OK
phyloP_hg38.npd.idx (GRCh38) OK
--------------------------------------------------------------------------------
--
Peak memory usage: 52.3 MB
Time: 00:04:12.2
```

ファイルへのアノテーションの付加

1. VCFファイルをまだ生成していない場合は、以下のコマンドを使用して、VCFファイルをダウンロードします。

   ```
   curl -O
   https://raw.githubusercontent.com/HelixGrind/DotNetMisc/master/TestFiles/H
   iSeq.10000.vcf.gz
   ```

   IAE は、圧縮されていない VCF ファイルと、bgzip 圧縮された VCF ファイルをサポートします。標準の gzip で圧縮された VCF ファイルはサポートされていません。

2. ファイルにアノテーションを付加するには、以下のコマンドを入力します：

   ```
   /opt/edico/share/nirvana/Nirvana -c ~/Data/Cache/GRCh38/Both \ -r
   ~/Data/References/Homo_sapiens.GRCh38.Nirvana.dat \ --sd
   ~/Data/SupplementaryAnnotation/GRCh38 -i HiSeq.10000.vcf.gz -o HiSeq.10000
   ```

使用可能なコマンドラインオプションは以下のとおりです：

オプション	値	例	説明
-c	ディレクトリ	~/Data/Cache/GRCh38/Both	キャッシュディレクトリ
-r	ディレクトリ	~/Data/References/Homo_sapiens.GRCh38.Nirvana.dat	リファレンスディレクトリ
--sd	ディレクトリ	~/Data/SupplementaryAnnotation/GRCh38	補完的アノテーションディレクトリ

オプション	値	例	説明
-i	パス	HiSeq.10000.vcf.gz	インプットVCFパス
-o	接頭辞	HiSeq.10000	アウトプットパスの接頭辞

前述の例を使って、IAEは、以下のアウトプット`HiSeq.10000.json.gz`を生成します。

```
------------------------------------------------------------------------
--
Nirvana (c) 2020 Illumina, Inc.
Stromberg, Roy, Lajugie, Jiang, Li, and Kang 3.9.1-0-gc823805
------------------------------------------------------------------------
--
Initialization Time Positions/s
------------------------------------------------------------------------
--
Cache 00:00:01.9
SA Position Scan 00:00:00.4 23,867
Reference Preload Annotation Variants/s
------------------------------------------------------------------------
--
chr1 00:00:00.4 00:00:03.7 2,651
Summary Time Percent
------------------------------------------------------------------------
--
Initialization 00:00:02.3 25.7 %
Preload 00:00:00.4 5.4 %
Annotation 00:00:03.7 41.5 %
Peak memory usage: 1.284 GB
Time: 00:00:08.0
```

JSON出力ファイル

IAEはJSON形式の出力ファイルを1つ作成します。このファイルには、以下の3つのセクションが含まれます：

セクション	内容
Header	構成、データソースのバージョン、サンプル名。
Positions	バリアントレベルのアノテーション。
Genes	遺伝子レベルのアノテーション。

前述のセクションで指定されたサンプルコマンドを使用したアウトプットを以下に示します。以下の各セクションには、情報の一部しか含まれていません。出力ファイルには、さらに多くの情報が含まれています。

Header

以下に、Headerセクションの例を示します。


```
"header": {
"annotator": "Nirvana 3.9.0",
"creationTime": "2020-06-03 08:05:06",
"genomeAssembly": "GRCh38",
"schemaVersion": 6,
"dataVersion": "91.26.57",
"dataSources": [
{
"name": "VEP",
"version": "91",
"description": "BothRefSeqAndEnsembl",
"releaseDate": "2018-03-05"
},
{
"name": "ClinVar",
"version": "20200302",
"description": "A freely accessible, public archive of reports of the
relationships among human variations and phenotypes, with supporting
evidence",
"releaseDate": "2020-03-02"
},
{
"name": "dbSNP",
"version": "153",
"description": "Identifiers for observed variants",
"releaseDate": "2019-07-22"
},
{
"name": "gnomAD",
"version": "2.1",
"description": "gnomAD allele frequency data remapped to GRCh38 with
CrossMap by Ensembl",
"releaseDate": "2019-03-25"
},
```

```
{
"name": "PrimateAI",
"version": "0.2",
"description": "PrimateAI percentile scores.",
"releaseDate": "2018-11-07"
},
{
"name": "OMIM",
"version": "20200409",
"description": "An Online Catalog of Human Genes and Genetic Disorders",
"releaseDate": "2020-04-09"
}
],
"samples": [
"NA12878"
]
},
```

Positions

1つのpositionは、VCFファイルの1行を表します。各positionには、samplesセクションとvariantsセクションが含まれます。ここに示されるリファレンスとオルタネティブアリルは、VCFの内容と正確に一致します。

samplesセクションには、遺伝型のようなサンプル固有の情報が、VCFと前述のJSONヘッダーに現れる順番で、表示されます。

variantsセクションには、VCF行にある各オルタネティブアリルのアノテーションが表示されます。これには、外部データソースからのアリル固有のアノテーションと転写産物レベルのアノテーションが含まれます。ここに示されるリファレンスとオルタネティブアリルは、最短の表現で示されます。例えば、パディングに使われた塩基が削除され、バリアントが左揃えされています。

```
"positions": [
{
"chromosome": "chr1",
"position": 1043248,
"refAllele": "C",
"altAlleles": [
"T"
],
"quality": 441.42,
"filters": [
```

```
"PASS"
],
"strandBias": -425.94,
"cytogeneticBand": "1p36.33",
"samples": [
{
"genotype": "0/1",
"variantFrequencies": [
0.537
],
"totalDepth": 54,
"genotypeQuality": 99,
"alleleDepths": [
25,
29
]
}
],
"variants": [
{
"vid": "1-1043248-C-T",
"chromosome": "chr1",
"begin": 1043248,
"end": 1043248,
"refAllele": "C",
"altAllele": "T",
"variantType": "SNV",
"hgvsg": "NC_000001.11:g.1043248C>T",
"phylopScore": 0.1,
"clinvar": [
{
"id": "RCV000872112.1",
"variationId": 263161,
"reviewStatus": "criteria provided, single submitter",
"alleleOrigins": [
"germline"
],
"refAllele": "C",
"altAllele": "T",
```

```
"phenotypes": [
"not provided"
],
"medGenIds": [
"CN517202"
],
"significance": [
"likely benign"
],
"lastUpdatedDate": "2019-12-17",
"pubMedIds": [
"28492532"
],
"isAlleleSpecific": true
},
{
"id": "VCV000263161.2",
"reviewStatus": "criteria provided, multiple submitters, no conflicts",
"significance": [
"likely benign"
],
"refAllele": "C",
"altAllele": "T",
"lastUpdatedDate": "2019-12-17",
"isAlleleSpecific": true
}
],
"dbsnp": [
"rs116586548"
],
"globalAllele": {
"globalMinorAllele": "T",
"globalMinorAlleleFrequency": 0.004393
},
"gnomad": {
"coverage": 38,
"allAf": 0.000681,
"allAn": 264462,
"allAc": 180,
```

```
      "allHc": 0,
      "afrAf": 0.006216,
      "afrAn": 23648,
      "afrAc": 147,
      "afrHc": 0,
      "amrAf": 0.000689,
      "amrAn": 33404,
      "amrAc": 23,
      "amrHc": 0,
      "easAf": 0,
      "easAn": 18830,
      "easAc": 0,
      "easHc": 0,
      "finAf": 0,
      "finAn": 22870,
      "finAc": 0,
      "finHc": 0,
      "nfeAf": 5e-05,
      "nfeAn": 120576,
      "nfeAc": 6,
      "nfeHc": 0,
      "asjAf": 0.000304,
      "asjAn": 9882,
      "asjAc": 3,
      "asjHc": 0,
      "sasAf": 0,
      "sasAn": 28456,
      "sasAc": 0,
      "sasHc": 0,
      "othAf": 0.000147,
      "othAn": 6796,
      "othAc": 1,
      "othHc": 0,
      "maleAf": 0.000564,
      "maleAn": 143614,
      "maleAc": 81,
      "maleHc": 0,
      "femaleAf": 0.000819,
      "femaleAn": 120848,
```

```
"femaleAc": 99,
"femaleHc": 0,
"controlsAllAf": 0.000626,
"controlsAllAn": 113456,
"controlsAllAc": 71
},
"oneKg": {
"allAf": 0.004393,
"afrAf": 0.016641,
"amrAf": 0,
"easAf": 0,
"eurAf": 0,
"sasAf": 0,
"allAn": 5008,
"afrAn": 1322,
"amrAn": 694,
"easAn": 1008,
"eurAn": 1006,
"sasAn": 978,
"allAc": 22,
"afrAc": 22,
"amrAc": 0,
"easAc": 0,
"eurAc": 0,
"sasAc": 0
},
"primateAI": [
{
"hgnc": "AGRN",
"scorePercentile": 0.12
}
],
"revel": {
"score": 0.136
},
"spliceAI": [
{
"hgnc": "AGRN",
"acceptorGainScore": 0.1,
```

```
"acceptorGainDistance": 23,
"acceptorLossScore": 0,
"acceptorLossDistance": -9,
"donorGainScore": 0,
"donorGainDistance": -5,
"donorLossScore": 0,
"donorLossDistance": 16
}
],
"topmed": {
"allAf": 0.002055,
"allAn": 125568,
"allAc": 258,
"allHc": 1
},
"transcripts": [
{
"transcript": "ENST00000379370.6",
"source": "Ensembl",
"bioType": "protein_coding",
"codons": "cCg/cTg",
"aminoAcids": "P/L",
"cdnaPos": "1444",
"cdsPos": "1394",
"exons": "8/36",
"proteinPos": "465",
"geneId": "ENSG00000188157",
"hgnc": "AGRN",
"consequence": [
"missense_variant"
],
"hgvsc": "ENST00000379370.6:c.1394C>T",
"hgvsp": "ENSP00000368678.2:p.(Pro465Leu)",
"isCanonical": true,
"polyPhenScore": 0.065,
"polyPhenPrediction": "benign",
"proteinId": "ENSP00000368678.2",
"siftScore": 0.05,
"siftPrediction": "tolerated"
```

```
},
{
"transcript": "NM_198576.3",
"source": "RefSeq",
"bioType": "protein_coding",
"codons": "cCg/cTg",
"aminoAcids": "P/L",
"cdnaPos": "1444",
"cdsPos": "1394",
"exons": "8/36",
"proteinPos": "465",
"geneId": "375790",
"hgnc": "AGRN",
"consequence": [
"missense_variant"
],
"hgvsc": "NM_198576.3:c.1394C>T",
"hgvsp": "NP_940978.2:p.(Pro465Leu)",
"isCanonical": true,
"polyPhenScore": 0.065,
"polyPhenPrediction": "benign",
"proteinId": "NP_940978.2",
"siftScore": 0.05,
"siftPrediction": "tolerated"
}
]
}
]
}
],
```


Genes

Positionsセクションの転写産物で参照される各遺伝子について、Genesセクションに一致するエントリがあります。以下に、gnomAD、ClinGen Dosage Sensitivity Map、およびOMIMからの遺伝子レベルのアノテーションの例を示します。


```
"genes": [
{
```

```
"name": "AGRN",
"gnomAD": {
"pLi": 5.47e-07,
"pRec": 1,
"pNull": 1.41e-12,
"synZ": -3.96,
"misZ": 0.226,
"loeuf": 0.435
},
"clingenDosageSensitivityMap": {
"haploinsufficiency": "gene associated with autosomal recessive
phenotype",
"triplosensitivity": "no evidence to suggest that dosage sensitivity is
associated with clinical phenotype"
},
"omim": [
{
"mimNumber": 103320,
"geneName": "Agrin",
"description": "The AGRN gene encodes agrin, a large and ubiquitous
proteoglycan with multiple isoforms that have diverse functions in
different tissues. Agrin was originally identified as an essential neural
regulator that induces the aggregation of acetylcholine receptors (AChRs)
and other postsynaptic proteins on muscle fibers and is crucial for the
formation and maintenance of the neuromuscular junction (NMJ) (Campanelli
et al., 1991; Burgess et al., 1999; summary by Maselli et al., 2012).",
"phenotypes": [
{
"mimNumber": 615120,
"phenotype": "Myasthenic syndrome, congenital, 8, with pre- and
postsynaptic defects",
"description": "Congenital myasthenic syndromes are genetic disorders of
the neuromuscular junction (NMJ) that are classified by the site of the
transmission defect: presynaptic, synaptic, and postsynaptic. CMS8 is an
autosomal recessive disorder characterized by prominent defects of both
the pre- and postsynaptic regions. Affected individuals have onset of
muscle weakness in early childhood; the severity of the weakness and
muscles affected is variable (summary by Maselli et al., 2012).\n\nFor a
discussion of genetic heterogeneity of CMS, see CMS1A.",
```

```
"mapping": "molecular basis of the disorder is known",
"inheritances": [
"Autosomal recessive"
]
}
]
}
]
}
]
}
```

DRAGEN ORA圧縮と展開

FASTQファイルの圧縮には、DRAGEN ORA圧縮も使用可能です。*.ora圧縮は、*.gz圧縮に置き換わります。DRAGEN ORAは、イルミナシーケンスシステムで生成された全FASTQファイルをサポートします。ORA形式を使用すると、可逆圧縮を保証するために、圧縮/展開サイクル後もFASTQコンテンツのmd5チェックサムが保持されます。

DRAGEN ORA圧縮には、別途、ライセンスが必要です。fastq.oraファイルの展開とDRAGENマッピング/アライメントへの統合にはライセンスは必要ありません。使用しているDRAGENサーバーがネットワークに接続されている場合、DRAGEN v3.8以降のインストール後、ORA圧縮を使用できます。DRAGENサーバーがオフラインの場合は、イルミナのカスタマーサービスへご連絡ください。

NovaSeq 6000、NextSeq 1000、またはNextSeq 2000シーケンスシステムで生成したヒトデータについては、圧縮比が*.fastq.gzの4～6倍になると予想されます。圧縮ファイルの拡張子には*.fastq.oraが使用されます。

.fastq filesや.fastq.gzファイルを*.fastq.oraに圧縮できます。また、*.fastq.oraを展開して*.fastq.gzにすることもできます。1つの入力ファイル、または複数のファイルを圧縮できます。複数のファイルを圧縮するときは、コマンドラインか、BCL Convert BaseSpace Sequence Hub AppまたはDRAGEN BCL変換で生成された*.fastq-list.csvでファイルのリストを指定します。

コマンドラインオプション

以下のコマンド例には、DRAGEN ORAの必須圧縮オプションが含まれます。

```
dragen --enable-map-align false --ora-input <FILE> --enable-ora true --
ora-reference <...> --output-directory <...>
```

以下のコマンド例には、ORAの必須展開オプションが含まれます。

```
dragen --enable-map-align false --ora-input <FILE> --enable-ora true --
ora-decompress true --ora-reference <...> --output-directory <...>
```

以下のコマンド例は、ORA圧縮ファイルのファイル情報サマリーを出力します。圧縮や展開は行われません。

```
dragen --enable-map-align false --ora-input <FILE> --enable-ora=true --
ora-print-file-info
```

DRAGEN ORA圧縮および展開の実行に使用可能なコマンドラインオプションは以下のとおりです。

オプション	必須かどうか	説明
--enable-map-align	必須	`false`に設定します。
--enable-ora	必須	FASTQファイルの圧縮と展開を有効化するには、trueに設定します。展開は、`--ora-decompress`オプションを使って有効化する必要があります。
--ora-reference	必須	圧縮リファレンスとインデックスファイルを含むディレクトリへのパス。
--ora-input	必須	圧縮または展開に使用する入力ファイルを指定します。
--ora-input2	必須ではない	ペア圧縮を行う2つめのファイルリストを指定します。ファイルの数は、`--ora-input`と同じでなければなりません。
--ora-decompress	必須ではない	圧縮モードを有効化するには、`true`に設定します。初期設定値はfalseです。
--force	必須ではない	圧縮済みのファイルが既に存在する場合でも、アウトプットディレクトリに圧縮します。既存の圧縮済みファイルは上書きされます。
--ora-threads-per-file <#>-	必須ではない	各FASTQ入力ファイルの圧縮に使用するCPUスレッドの数を人為的にコントロールします。初期設定値は8です。
--ora-parallel-files <#>	必須ではない	同時処理されるFASTQ入力ファイルの数をマニュアルでコントロールします。初期設定値は4です。
--ora-use-hw	必須ではない	ハードウェアアクセラレーションを有効化する場合はtrue、無効化する場合はfalseに設定します。初期設定値は`true`です。
--ora-print-file-info	必須ではない	ORA圧縮ファイルのファイル情報サマリーを出力します。

出力された圧縮/展開ファイルの保存先ディレクトリを指定するには、--output-directoryオプションを使用します。

圧縮に--ora-inputを使用する代わりに、BCL Convert AppまたはDRAGEN BCL変換で生成されたfastq-list.csvを使用することもできます。

- fastq-list内の全ファイルを圧縮するには、--fastq-list <fastq_list csv> --fastq-list-all-samples trueを使用します。
- 特定のサンプルのファイルだけを圧縮するには、--fastq-list <fatsq_list csv> --fastq-list-sample-idを使用します。

ペア圧縮

--ora-inputと--ora-input2を使用した場合、圧縮はペアモードで行われます。ペア圧縮では、--ora-inputリストのn番目のファイルが、--ora-input2のn番目のファイルとともに圧縮されます。ファイルは両方とも、1つのORA出力ファイルにインターリーブされます。これらのオプションを使って、対になるファイルを一緒に圧縮することができます。これにより、圧縮効率が最大10%向上します。ペアデータを含むORAファイルを展開すると、1つのファイルが、自動的に2つのファイルへ展開されます。DRAGENマッパーでインターリーブされたペアデータを含むORAファイルをマッピングするには、--interleavedオプションを使用します。

ORAリファレンス

ORAファイルを圧縮または展開するには、ORAリファレンスファイルを提供し、ORAリファレンスディレクトリを指定する必要があります。ORAリファレンスファイルは、Illumina DRAGEN Bio-IT Platformサポートサイトページからダウンロードできます。

ORAリファレンスディレクトリを指定するには、以下のように操作します。

1. DRAGEN Bio-It Platformのサポートサイトから、lenadata-1.tar.gzをダウンロードします。
2. このファイルを、リファレンスディレクトリを配置したい場所に移動し、以下のコマンドを入力して、ファイルの中身を抽出します。

   ```
   tar -xzvf lenadata-1.tar.gz
   ```

3. --ora-referenceコマンドラインオプションに、抽出した/lenadataフォルダーパスを設定します。

使用モード

DRAGEN ORA圧縮/展開のFPGAアクセラレーションを使用するには、--ora-use-hwをtrueに設定します。--ora-use-hwをfalseに設定した場合、オンサイトシステムの使用時に、FPGAを使用する他のプロセスと同時にDRAGEN ORA圧縮/展開を起動することができます。クラウドでは、DRAGEN ORA圧縮/展開と、DRAGENデータ解析の同時実行はサポートされていません。

ハードウェアアクセラレーションによる圧縮と展開

gzip圧縮は、バイオインフォマティクスでは広く使われています。FASTQファイルがgzipされていることはよくあり、BAM形式自体がgzipの特別なバージョンです。そのため、DRAGEN BioITプロセッサーは、gzipされたデータの圧縮と展開を加速するためのハードウェアサポートを提供しています。入力ファイルがgzipされていた場合、DRAGENはそれを検出し、自動的に展開します。出力がBAMファイルの場合、ファイルは自動的に圧縮されます。

DRAGENには、任意のファイルを圧縮または展開できるようにするための、スタンドアロンのコマンドラインユーティリティが用意されています。これらのユーティリティは、Linuxのgzipやgunzipコマンドに似ていますが、名前は*dzip*と*dunzip*（dragen zipとdragen unzip）です。どちらのユーティリティも1つのファイルを入力として受け付け、1つの出力ファイルを作成することができます。また、必要に応じて、ファイル拡張子.gzを削除または追加できます。例えば：

```
dzip file1        # produces output file file1.gz
dunzip file2.gz   # produces output file file2
```

現時点では、*dzip*と*dunzip*には、以下のような制約およびgzip/gunzipとの違いがあります：

- 1回の呼び出しで処理できるファイルは1つだけです。追加のファイル名（ワイルドカード文字「*」で指定されたものを含む）は無視されます。
- DRAGENホストソフトウェアと同時に実行できません。
- gzipとgunzipの持つコマンドラインオプション（--recursive、--fast、--best、--stdoutなど）をサポートしていません。

使用状況レポートの作成

インストールプロセスの一環として、デーモン（dragen_licd）が作成されます（または、一旦停止し、その後再開されます）。このバックグラウンドプロセスは、毎日、1日の終わりに自らをアクティブ化し、DRAGENホストソフトウェアの使用状況をイルミナサーバーにアップロードします。情報には、日付、稼働期間、サイズ（塩基の数）、各ランのステータス、使用したソフトウェアのバージョンが含まれます。

イルミナサーバーとの通信は、暗号化により保護されています。通信エラーが発生した場合、デーモンは翌朝まで再試行します。アップロードの失敗が続いた場合、翌日の晩、通信に成功するまで、引き続き再試行されます。つまり、営業時間中は、システムリソースをフルに利用することができ、いかなる形でも、このバックグラウンド処理によって業務が妨げられることはありません。

現在のライセンス使用状況を確認するには、*dragen_lic*コマンドを使用します。

使用状況レポートを生成するためサーバー要件は以下のとおりです：

- 300x生殖細胞系列シングルサンプルカバレッジに対して256 GBのRAMおよび2 TBのHDD。
- T/N解析カバレッジ。
- 6 TBおよび512 GBのRAM。

使用状況レポートには、以下の情報は含まれません：

- ジョイントジェノタイパー入力ファイルの数。
- gVCFジェノタイパーへのGATK gVCFインプット。
- ジョイントコールやgVCFジェノタイパーなど異なるコーラーからのミキシングgVCF。

トラブルシューティング

DRAGENシステムが応答していないと思われる場合、以下の操作を行います：

1. DRAGENシステムがハングしていないかどうかを判断するために、351 ページの 「システムがハングしているかどうかを判断するには」 の指示に従って操作します。
2. 351 ページの 「イルミナサポートへの診断データの送信」 に従って、ハング、またはクラッシュ後の診断情報を収集します。
3. 情報をすべて収集し終えたら、必要に応じて、351 ページの 「クラッシュまたはハング後に、システムをリセットするには」 の手順に従って、システムをリセットします。

システムがハングしているかどうかを判断するには

DRAGENシステムには、システムのハングをモニタリングするウォッチドッグ機能があります。ランの時間が普通より長いと感じられる場合、ウォッチドッグ機能がハングを検出していない可能性があります。以下を試してみてください。

- *top*コマンドを実行して、アクティブなDRAGENプロセスを見つけます。ランが正常に行われていれば、ランのCPU消費量が100%を超えていることが表示されるはずです。消費が100%以下の場合、システムはハングしている可能性があります。
- 出力BAM/SAMファイルのディレクトリで、*du -s*コマンドを実行します。通常のランの間、このディレクトリは、中間出力データ（ソートが有効の場合）またはBAM/SAMデータで大きくなり続けているはずです。

イルミナサポートへの診断データの送信

イルミナは、DRAGENシステムに関するフィードバックを歓迎いたします。これにはシステムの不具合に関するレポートも含まれます。クラッシュ、ハング、ウォッチドッグの不具合などが発生した際には、以下のように、*sosreport*コマンドを実行して、診断情報と構成情報を収集します：

```
sudo sosreport --batch --tmp-dir /staging/tmp
```

このコマンドの実行には数分かかります。また、このコマンドは、`/staging/tmp`に保存した診断情報の場所をレポートします。イルミナのテクニカルサポートにチケットを送信するときにはこのレポートを添付してください。

クラッシュまたはハング後に、システムをリセットするには

DRAGENシステムがクラッシュまたはハングした場合、*dragen_reset*ユーティリティを実行して、ハードウェアとソフトウェアを再初期化する必要があります。ホストソフトウェアは、予期せぬ状態を検出したときに必ず、このユーティリティを自動的に実行します。この場合、ホストソフトウェアから以下のメッセージが表示されます：

```
Running dragen_reset to reset DRAGEN Bio-IT processor and software
```

ソフトウェアがハングした場合は、351 ページの 「イルミナサポートへの診断データの送信」 の手順に従って、診断情報を収集してから、以下のように*dragen_reset*を実行してください：

```
/opt/edico/bin/dragen_reset
```

*dragen_reset*を実行したときには必ず、リファレンスゲノムをDRAGENボードに再ロードする必要があります。このリファレンスは、次回の実行時に、ホストソフトウェアにより自動的に再ロードされます。

コマンドラインオプションリファレンス

ここでは、DRAGENの全コマンドラインオプションについて、構成ファイルで使用される名前、名前に対応するコマンドラインオプション、説明、値の範囲などの情報を提供します。

一般的なソフトウェアオプション

以下のオプションは、構成ファイルの初期設定セクションで使用されるものです。初期設定セクションは、構成ファイルの先頭にありますが、[Aligner]のようなセクション名を持ちません。一部の必須ファイルは、コマンドラインで指定しなければならず、構成ファイルには登場しません。

名前	説明	対応するコマンドラインオプション	値
alt-aware	ALTリフトオーバーがハッシュテーブルで使用されている場合、このオプションは、ALTコンティグの特別な処理を有効にします。リファレンスが、リフトオーバーを使って構築されている場合、このオプションは初期設定で有効にされています。	`--alt-aware`	• true • false
append-read-index-to-name	初期設定では、ペアのメイトの両端に同じ名前が付けられます。trueに設定されている場合、2つの末端に/1と/2が付加されます。	`--append-read-index-to-name`	• true • false
bam-input	DRAGENバリアントコーラーへのインプットに使用するアライメント済みBAMファイルを指定します。	`-b, --bam-input`	
bcl-conversion-only	FASTQ形式へのIllumina BCLconversionを実行します。	`--bcl-conversion-only`	• true • false
bcl-input-directory	BCL変換に使用するBCLディレクトリをインプットします。	`--bcl-input-directory`	
bcl-only-lane	BCLインプットでは、指定されたレーン番号のみ変換されます。初期設定では、全レーンが変換されます。	`--bcl-only-lane`	1〜8

名前	説明	対応するコマンドライン オプション	値
sample-sheet	BCLインプットでは、SampleSheet.csvファイルへのパスを設定します。初期設定では、BCLのrootディレクトリです。	--sample-sheet	
strict-mode	BCLインプットでは、所在のわからないファイルが1つでもあった場合に、データ解析をキャンセルします。初期設定値はfalseです。	--strict-mode	• true • false
first-tile-only	BCL変換中、各レーンの先頭タイルのみ変換します。テストやデバッグに使用します。	--first-tile-only	• true • false
run-info	RunInfo.xmlファイルへのパスを設定します。初期設定は<flow cell>/RunInfo.xmlです。	--run-info	
bcl-sampleproject-subdirectories	BCL変換では、サンプルシートのSample_Project列に基づいて、サブディレクトリにアウトプットします。	--bcl-sampleproject-subdirectories	
no-lane-splitting	レーンによるFASTQ出力ファイルの分割を無効にします。初期設定値はfalseです。	--no-lane-splitting	• true • false
bcl-only-matched-reads	マッピングされていないリードを、Undeterminedとマークされたファイルにアウトプットするかどうかを指定します。初期設定値はfalseです。	bcl-only-matched-reads	• true • false
bcl-use-hw	falseに設定した場合、BCL変換中、DRAGEN FPGAアクセラレーションは使用されません。初期設定値はtrueです。	--bcl-use-hw	• true • false

名前	説明	対応するコマンドライン オプション	値
bcl-num-parallel-tiles	同時に処理するタイルの数を指定します。初期設定値は動的に決定されます。	`--bcl-num-parallel-tiles`	• 1〜<nproc>
bcl-num-conversion-threads	1タイルあたりの変換スレッド数を指定します。初期設定値は動的に決定されます。	`--bcl-num-conversion-threads`	• 1〜<nproc>
bcl-num-compression-threads	アウトプット`fastq.gz`圧縮に使用するCPUスレッドの数を指定します。初期設定値は動的に決定されます。	`--bcl-num-compression-threads`	• 1〜<nproc>
bcl-num-decompression-threads	BCL入力展開に使用するCPUスレッドの数を指定します。初期設定値は動的に決定されます。	`--bcl-num-decompression-threads`	• 1〜<nproc>
shared-thread-odirect-output	代替共有スレッドODIRECTファイル出力を使用します。初期設定値はfalseです。	`--shared-thread-odirect-output`	• true • false
build-hash-table	リファレンスハッシュテーブルを生成します。	`--build-hash-table`	• true • false
cram-input	DRAGENバリアントコーラーで使用されるCRAMファイル入力を指定します。	`--cram-input`	
dbsnp	バリアントアノテーションデータベースVCF（または*.vcf.gz）ファイルへのパスを設定します。	`--dbsnp`	
enable-auto-multifile	`*_001.{dbam,fastq}`ファイルの後続セグメントをインポートします。	`--enable-auto-multifile`	• true • false
enable-bam-indexing	BAIインデックスファイルの生成を有効にします。	`--enable-bam-indexing`	• true • false

名前	説明	対応するコマンドライン オプション	値
enable-cram-indexing	CRAIインデックスファイルの生成を有効にします。	`--enable-cram-indexing`	• true • false
enable-cnv	コピー数バリアント(CNV)を有効にします。	`--enable-cnv`	• true • false
enable-duplicate-marking	重複するアウトプットアライメントレコードのフラグ付けを有効にします。	`--enable-duplicate-marking`	• true • false
enable-map-align-output	マッピング/アライメントステージの出力の保存を有効にします。マッピング/アライメントのみを実行する場合、初期設定値はtrueです。バリアントコーラーを実行する場合、初期設定値はfalseです。	`--enable-map-align-output`	• true • false
enable-methylation-calling	メチル化に関連するタグを自動的に追加し、メチル化プロトコール用にBAMを1つアウトプットします。	`--enable-methylation-calling`	• true • false
enable-sampling	マッパー/アライナー経由でサンプルを実行し、ペアエンドパラメーターを自動検出します。	`--enable-sampling`	• true • false
enable-sort	マッピング/アライメント後のソーティングを有効にします。	`--enable-sort`	• true • false
enable-variant-caller	バリアントコーラーを有効にします。	`--enable-variant-caller`	• true • false
enable-variant-deduplication	バリアント重複除去を有効にします。初期設定値はfalseです。	`--enable-variant-deduplication`	• true • false
enable-vcf-compression	VCF出力ファイルの圧縮を有効にします。初期設定値はtrueです。	`--enable-vcf-compression`	• true • false
enable-vcf-indexing	出力VCF/gVCFに加えて、*.tbiインデックスファイルをアウトプットします。初期設定はtrueです。	`--enable-vcf-indexing`	• true • false

名前	説明	対応するコマンドライン オプション	値
fastq-file1	DRAGENパイプラインへインプットするFASTQファイルを指定します。gzipされた形式も使用できます。	`-1, --fastq-file1`	
fastq-file2	インプットするペアエンドリードの2つめのFASTQファイルを指定します。	`-2, --fastq-file2`	
fastq-list	処理するFASTQファイルのリストを含むCSVファイルを指定します。	`--fastq-list`	
fastq-list-sample-id	RGSMエントリーが、指定された`fastq-list.csv`インプットのSample IDパラメーターと一致する場合、このエントリーを処理します。	`--fastq-list-sample-id`	• true • false
fastq-list-all-samples	RGSM値に関係なく、すべてのサンプルの一括処理を無効にします。	`--fastq-list-all-samples`	• true • false
fastq-n-quality	N塩基にアウトプットするベースコールのクオリティを指定します。全アウトプットN塩基の`fastq-n-quality`に自動的に追加されます。	`--fastq-n-quality`	• 0～255
fastq-offset	FASTQクオリティのオフセット値を設定します。	`--fastq-offset`	• 33 • 64
filter-flags-from-output	flagsフィールドに存在する値に設定された任意のビット数で、アウトプットアライメントをフィルターします。16進数値と10進数値を使用できます。	`--filter-flags-from-output`	
force	既存の出力ファイルを強制的に上書きします。	`-f`	

名前	説明	対応するコマンドライン オプション	値
force-load-reference	DRAGENパイプラインを開始する前に、リファレンスとハッシュテーブルを強制的にローディングします。	`-l`	
generate-md-tags	アライメント出力レコードを使用して、MDタグを生成します。初期設定値はfalseです。	`--generate-md-tags`	• true • false
generate-sa-tags	キメラアライメントまたは補足アライメントを有するレコードに対し、SA:Zタグを生成します。	`--generate-sa-tags`	• true • false
generate-zs-tags	アライメント出力レコードに対し、ZSタグを生成します。初期設定値はfalseです。	`--generate-sz-tags`	• true • false
ht-alt-liftover	リファレンスにある代替コンティグのSAM形式のリフトオーバーファイル。	`--ht-alt-liftover`	
ht-mask-bed	塩基のマスキングに使用するBEDファイルを指定します。	`--ht-mask-bed`	
ht-allow-mask-and-liftover	ht-alt-liftoverとht-mask-bedの両方を指定して、ハッシュテーブルビルダーを実行できるようにします。初期設定はfalseです。	`--ht-allow-mask-and-liftover`	• true • false
ht-alt-aware-validate	ALTコンティグを使ってhg19またはhg38からハッシュテーブルを構築するときに、マスキング用BEDファイル、またはリフトオーバーファイルに対する要件を無効にします。	`--ht-alt-aware-validate`	• true • false
ht-build-rna-hashtable	RNAハッシュテーブルの生成を有効にします。初期設定値はfalseです。	`--ht-build-rna-hashtable`	• true • false
ht-cost-coeff-seed-freq	拡張シード頻度のコスト係数を設定します。	`--ht-cost-coeff-seed-freq`	

名前	説明	対応するコマンドライン オプション	値
ht-cost-coeff-seed-len	拡張シード長のコスト係数を設定します。	`--ht-cost-coeff-seed-len`	
ht-cost-penalty-incr	シードを1段階、拡張する際のコストペナルティを設定します。	`--ht-cost-penalty-incr`	
ht-cost-penalty	シードを任意の数の塩基分、拡張する際のコストペナルティを設定します。	`--ht-cost-penalty`	
ht-decoys	デコイファイルへのパスを指定します。	`--ht-decoys`	
ht-max-dec-factor	シードの間引きをする際の、最大間引き係数を設定します。	`--ht-max-dec-factor`	
ht-max-ext-incr	シードを1ステップ拡張する際の、最大塩基数を設定します。	`--ht-max-ext-incr`	
ht-max-ext-seed-len	拡張シード長の最大値を設定します。	`--ht-max-ext-seed-len`	
ht-max-seed-freq	拡張を試行した後のシードマッチの最大許容頻度を設定します。	`--ht-max-seed-freq`	• 1～256
ht-max-table-chunks	一度に最大～1 GBスレッドテーブルチャンクをメモリーに指定します。	`--ht-max-table-chunks`	
ht-mem-limit	メモリーの上限(ハッシュテーブル+リファレンス)をKB、MB、またはGB単位で指定します。	`--ht-mem-limit`	
ht-methylated	C- >TおよびG->A変換されたリファレンスハッシュテーブルを自動生成します。	`--ht-methylated`	• true • false
ht-num-threads	ハッシュテーブルを構築する際の、ワーカーCPUスレッドの最大値を設定します。	`--ht-num-threads`	

名前	説明	対応するコマンドライン オプション	値
ht-rand-hit-extend	頻度レコードの各EXTENDレコードにランダムヒットを含めます。	`--ht-rand-hit-extend`	
ht-rand-hit-hifreq	各HIFREQレコードにランダムヒットを含めます。	`--ht-rand-hit-hifreq`	
ht-ref-seed-interval	リファレンスシード1つあたりの位置数を指定します。	`--ht-ref-seed-interval`	
ht-reference	ハッシュテーブルを構築するための、FASTA形式のリファレンスファイルです。	`--ht-reference`	
ht-seed-len	ハッシュテーブルを格納するシード長の初期値を設定します。	`--ht-seed-len`	
ht-size	ハッシュテーブルのサイズをKB、MB、またはGB単位で指定します。	`--ht-size`	
ht-soft-seed-freq-cap	間引きをする際の、ソフトシード頻度の上限を指定します。	`--ht-soft-seed-freq-cap`	
ht-suppress-decoys	ハッシュテーブルを構築するときに、デコイファイルが使用されないようにします。	`--ht-suppress-decoys`	
ht-target-seed-freq	シード拡張の際のシード頻度の目標値を設定します。	`--ht-target-seed-freq`	
input-qname-suffix-delimiter	append-read-index-toname、およびBAMインプットとマッチするペア名の検出に使用する区切り文字をコントロールします。	`--input-qname-suffix-delimiter`	• / • . • :
interleaved	1つのFASTQ内のインターリーブされたペアエンドリードを指定します。	`-i`	
intermediate-results-dir	中間結果の格納先ディレクトリ(ソートパーティションなど)を指定します。	`--intermediate-results-dir`	

名前	説明	対応するコマンドライン オプション	値
lic-no-print	実行の最後に、ライセンスのステータスメッセージが表示されないようにします。	`--lic-no-print`	• true • false
methylation-generate-cytosine-report	ゲノムワイドなシトシンメチル化レポートを生成します。	`--methylation-generate-cytosine-report`	• true • false
methylation-generate-mbias-report	システムサイクル1回あたりのメチル化バイアスレポートを生成します。	`--methylation-generate-mbias-report`	• true • false
methylation-TAPS	インプットアッセイがTAPSにより生成された場合、このオプションはtrueに設定されます。	`--methylation-TAPS`	• true • false
methylation-match-bismark	trueの場合、このオプションはバグも含め、bismarkタグと正確に一致します。	`--methylation-match-bismark`	• true • false
methylation-protocol	メチル化解析のライブラリープロトコールを記述します。	`--methylation-protocol`	• none • directional • nondirectional • directional-complement
num-threads	使用するプロセッサースレッド数を指定します。	`-n, --num-threads`	
output-directory	アウトプットディレクトリを指定します。	`--output-directory`	
output-file-prefix	パイプラインで生成される全ファイルで使用する出力ファイル名の接頭辞。	`--output-file-prefix`	

名前	説明	対応するコマンドライン オプション	値
output-format	マッピング/アライメントステージの出力ファイルの形式を表します。次の値が有効です: • BAM（初期設定） • CRAM • SAM • DBAM（独自のバイナリー形式）	`--output-format`	• BAM • CRAM • SAM • DBAM
pair-by-name	ペアエンドメイトが一緒に処理されるように、BAMインプットレコードの順序を入れ替えます。	`--pair-by-name`	
pair-suffix-delimiter	接尾辞の区切り文字を変更します。	`--pair-suffix-delimiter`	• / • . • :
preserve-bqsr-tags	インプットBAMファイルのBIフラグとBDフラグを保持するかどうかを決定します。これは、ハードクリッピングに問題を引き起こすことがあります。	`--preserve-bqsr-tags`	• true • false
preserve-map-align-order	入力ファイル中のリードの順番をそのまま保持した出力ファイルを作成します。	`--preserve-map-align-order`	• true • false
qc-coverage-region-1	BEDファイル1を使用して、カバレッジ領域レポートを生成します。	`--qc-coverage-region-1`	
qc-coverage-region-2	BEDファイル2を使用して、カバレッジ領域レポートを生成します。	`--qc-coverage-region-2`	
qc-coverage-region-3	BEDファイル3を使用して、カバレッジ領域レポートを生成します。	`--qc-coverage-region-3`	
qc-coverage-reports-1	qc-coverage-region-1に対して要求されたレポートのタイプを表します。	`--qc-coverage-reports-1`	• full_res • cov_report

名前	説明	対応するコマンドライン オプション	値
qc-coverage-reports-2	qc-coverage-region-2に対して要求されたレポートのタイプを表します。	`--qc-coverage-reports-2`	• full_res • cov_report
qc-coverage-reports-3	qc-coverage-region-3に対して要求されたレポートのタイプを表します。	`--qc-coverage-reports-3`	• full_res • cov_report
ref-dir	リファレンスハッシュテーブルを含むディレクトリを指定します。リファレンスがDRAGENカードにロードされていない場合、このオプションは自動的にリファレンスをロードします。	`-r, --ref-dir`	
ref-sequence-filter	アウトプットは、リファレンスシーケンスへのマッピングの読み込みだけを行います。	`--ref-sequence-filter`	
remove-duplicates	trueの場合、このオプションは、重複するアライメントレコードにフラグを付けて済ませる代わりに、該当するレコードを削除します。		• true • false
RGCN	リードグループのシーケンシングセンター名を指定します。	`--RGCN`	
RGCN-tumor	腫瘍グループのリードグループのシーケンシングセンター名を指定します。	`--RGCN-tumor`	
RGDS	リードグループの説明を提供します。	`--RGDS`	
RGDS-tumor	腫瘍インプットについて、リードグループの説明を提供します。	`--RGDS-tumor`	
RGDT	リードグループの実行日を指定します。	`--RGDT`	
RGDT-tumor	腫瘍インプットについて、リードグループの実行日を指定します。	`--RGDT-tumor`	
RGID	リードグループIDを指定します。	`--RGID`	

名前	説明	対応するコマンドライン オプション	値
RGID-tumor	腫瘍インプットについて、リードグループIDを指定します。	`--RGID-tumor`	
RGLB	リードグループライブラリーを指定します。	`--RGLB`	
RGLB-tumor	腫瘍インプットについて、リードグループライブラリーを指定します。	`--RGLB-tumor`	
RGPI	リードグループが予測したインサートサイズを指定します。	`--RGPI`	
RGPI-tumor	腫瘍インプットについて、リードグループが予測したインサートサイズを指定します。	`--RGPI-tumor`	
RGPL	リードグループのシーケンシングテクノロジーを指定します。	`--RGPL`	
RGPL-tumor	腫瘍インプットについて、リードグループのシーケンシングテクノロジーを指定します。	`--RGPL-tumor`	
RGPU	リードグループのプラットフォームユニットを指定します。	`--RGPU`	
RGPU-tumor	腫瘍インプットについて、リードグループのプラットフォームユニットを指定します。	`--RGPU-tumor`	
RGSM	リードグループのサンプル名を指定します。	`--RGSM`	
RGSM-tumor	腫瘍グループについて、リードグループのサンプル名を指定します。	`--RGSM-tumor`	
sample-size	enable-samplingがtrueの場合にサンプリングするリードの数を指定します。	`--sample-size`	
sample-sex	サンプルの性を指定します。	`--sample-sex`	

名前	説明	対応するコマンドライン オプション	値
strip-input-qname-suffixes	インプットのQNAMEから、リードインデックス接尾辞（/1、/2など）を取り除くかどうかを決定します。falseに設定した場合、名前はそのまま保持されます。	`--strip-input-qname-suffixes`	• true • false
tumor-bam-input	体細胞モードで、DRAGENバリアントコーラーに使用するアライメント済みBAMファイルを指定します。	`--tumor-bam-input`	
tumor-cram-input	体細胞モードで、DRAGENバリアントコーラーに使用するアライメント済みCRAMファイルを指定します。	`--tumor-cram-input`	
tumor-fastq-list	FASTQファイルのリスト含むCSVファイルを、マッパー、アライナーおよび体細胞バリアントコーラーに入力します。	`--tumor-fastq-list`	
tumor-fastq-list-sample-id	tumor-fastq-listで指定されたFASTQファイルのリストのサンプルIDを指定します。	`--tumor-fastq-list-sample-id`	
tumor-fastq1	体細胞モードで、バリアントコーラーを使用して、DRAGENパイプラインのFASTQファイルを入力します。入力ファイルはgzipできます。	`--tumor-fastq1`	
tumor-fastq2	2つめのFASTQファイルをインプットします。リードは、体細胞モードで、バリアントコーラーを使用して、DRAGENパイプラインのtumorfastq1リードとペアにされます。入力ファイルはgzipできます。	`--tumor-fastq2`	
vd-eh-vcf	バリアント重複除去のためのExpansionHunter VCFファイルを入力します。入力ファイルはgzipできます。	`--vd-eh-vcf`	

名前	説明	対応するコマンドライン オプション	値
vd-output-match-log	重複除去中にマッチしたバリアントを記述したファイルを出力します。初期設定値はfalseです。	`--vd-output-match-log`	• true • false
vd-small-variant-vcf	バリアント重複除去のためのスモールバリアントVCFファイルを入力します。入力ファイルはgzipできます。	`--vd-small-variant-vcf`	
vd-sv-vcf	バリアント重複除去のための構造多型VCFファイルを入力します。入力ファイルはgzipできます。	`--vd-sv-vcf`	
verbose	DRAGENからのverboseアウトプットを有効にします。	`-v`	
version	バージョンを表示して、終了します。	`-V`	

マッパーオプション

以下のオプションは、構成ファイルの[Mapper]セクションで使用されます。以下のオプションの詳細については、65 ページの 「DNAマッピング」 を参照してください。

名前	説明	対応するコマンドライン オプション	値
ann-sj-max-indel	アノテーションされたスプライスジャンクション付近で予測されるIndel長の最大値を指定します。	`--Mapper.ann-sj-max-indel`	• 0～63
edit-chain-limit	edit-mode 1または2では、このオプションは、シード編集の対象となりうる、リード内のシードチェーン長の最大値を設定します。	`--Mapper.edit-chain-limit`	• edit-chain-limit >= 0

名前	説明	対応するコマンドライン オプション	値
edit-mode	シード編集が使用される時期をコントロールします。以下の値はそれぞれ異なる編集モードを表します: • 0：編集しない • 1：鎖長テスト • 2：ペアード鎖長テスト • 3：シード全編集	`--Mapper.edit-mode`	• 0～3
edit-read-len	edit-modeが1または2の場合、edit-seednumシード編集位置のリード長をコントロールします。	`--Mapper.edit-read-len`	• edit-read-len > 0
edit-seed-num	edit-modeが1または2の場合、編集を可能にするために要求される1リードあたりのシード数をコントロールします。	`--Mapper.edit-seed-num`	• edit-seed-num >= 0
enable-map-align	マッパー/アライナーでBAM入力ファイルの使用を可能にします。	`--enable-map-align`	• true • false
map-orientations	リードマッピングの方向をリファレンスゲノムのフォワード方向のみ、または逆相補方向のみに制限します。以下の値はそれぞれ異なる方向を表します（ペアエンドでは正常が必要です）: • 0：正常（ペアエンド入力は正常を使用する必要があります） • 1：逆相補方向 • 2：フォワード方向	`--Mapper.map-orientations`	• 0～2
max-intron-bases	レポートされたイントロン長の最大値を指定します。	`--Mapper.max-intron-bases`	
min-intron-bases	イントロンとしてレポートされたリファレンス欠失長の最小値を指定します。	`--Mapper.min-intron-bases`	
seed-density	ハッシュテーブルでクエリされたリードから得られた、要求されたシード密度をコントロールします。	`--Mapper.seed-density`	• 0 > seed-density > 1

アライナーのオプション

以下のオプションは、構成ファイルの[Aligner]セクションで使用されます。詳細については、68 ページの「DNAアライメント」 を参照してください。

名前	説明	対応するコマンドラインオプション	値
aln-min-score	MAPQのベースラインをレポートするために、許容可能なアライメントスコアの最小値を指定する符号付き整数。ローカルアライメントを使用する場合(global=0)、aln-min-scoreは、ホストソフトウェアにより、22 * match-scoreの式で計算されます。 グローバルアライメントを使用する場合(global=1)、`aln-min-score`は-1000000に設定されます。ホストソフトウェアによる計算をオーバーライドするには、構成ファイルに`aln-min-score`を設定します。	`--Aligner.aln-min-score`	• −2147483648 〜 2147483647
dedup-min-qual	重複除去のためにリードクオリティメトリクスを計算するための、塩基クオリティの最小値を指定します。	`--Aligner.dedup-min-qual`	• 0〜63
en-alt-hap-aln	キメラアライメントを補足としてアウトプットすることを許可します。	`--Aligner.en-alt-hap-aln`	• 0〜1
en-chimeric-aln	キメラアライメントを補足としてアウトプットすることを許可します。	`--Aligner.en-chimeric-aln`	• 0〜1
gap-ext-pen	ギャップの拡張に対するペナルティを指定します。	`--Aligner.gap-ext-pen`	• 0〜15
gap-open-pen	ギャップ(挿入または欠失)を開けることに対するペナルティを指定します。	`gap-open-pen`	• 0〜127

名前	説明	対応するコマンドライン オプション	値
global	アライメントが、リード内でエンドツーエンドかどうかをコントロールします。以下の値はそれぞれ異なるアライメントを表します: • 0：ローカルアライメント（Smith-Waterman） • 1：グローバルアライメント（Needleman-Wunsch）	`--Aligner.global`	• 0〜1
hard-clips	ハードクリッピングに対するアライメントを指定します。以下の値はそれぞれ異なるアライメントを表します: • ビット0は一次 • ビット1は補足 • ビット2は二次	`--Aligner.hard-clips`	• 3ビット
map-orientations	フォワードのみ、逆相補のみ、または任意のアライメントを受け付けるために方向を制約します。 以下の値はそれぞれ異なる方向を表します: • 0：任意 • 1：フォワードのみ • 2：逆相補のみ	`--Aligner.map-orientations`	• 0〜2
mapq-max	レポートされたMAPQの上限を指定します。初期設定値は60です。	`--Aligner.mapq-max`	• 0〜255

名前	説明	対応するコマンドライン オプション	値
mapq-strict-js	RNA固有です。0に設定すると、より高いMAPQ値が返信されるため、そのアライメントが少なくとも部分的に正しいという信頼度が表されます。1に設定すると、より低いMAPQ値が返信されるため、スプライスジャンクションの曖昧さが表されます。	`--mapq-strict-js`	• 0〜1
match-n-score	リードまたはリファレンス塩基がNであるマッチングのスコア増分を指定する符号付き整数。	`--Aligner.match-n-score`	• -16〜15
match-score	マッチするリファレンスヌクレオチドのスコア増分を指定します。	`--Aligner.match-score`	• global = 0のとき、match-score > 0 • global = 1のとき、match-score >= 0
max-rescues	リードペア1つあたりの最大レスキューアライメントを指定します。初期設定値は10です。	`--max-rescues`	• 0〜1023
min-score-coeff	リード塩基1つあたりの調整サイズを`aln-min-score`に設定します。	`--Aligner.min-score-coeff`	• -64〜63.999
mismatch-pen	ミスマッチのスコアペナルティを定義します。	`--Aligner.mismatch-pen`	• 0〜63

名前	説明	対応するコマンドライン オプション	値
no-unclip-score	1に設定されているとき、アライメントに寄与するクリップされていないときのボーナス(unclip-score)は、次のプロセスの前にアライメントスコアから除かれます。	--Aligner.no-unclip-score	• 0〜1
no-unpaired	ペアリードについては、適切にペアリングされたアライメントのみをレポートする必要があるかどうかを決定します。	--Aligner.no-unpaired	• 0〜1
pe-max-penalty	ペアリングされていないエンド、または距離の離れたエンドに対するペアリングスコアペナルティの最大値を指定します。	--Aligner.pe-max-penalty	• 0〜255
pe-orientation	予想されるペアエンドの方向を指定します。 以下の値はそれぞれ異なる方向を表します: • 0はFR（初期設定） • 1はRF • 2はFF	--Aligner.pe-orientation	• 0〜2
rescue-sigmas	レスキュースキャン範囲に使用される平均リード長からの逸脱を設定します。初期設定値は2.5です。	--Aligner.rescue-sigmas	
sec-aligns	レポートの対象となる、1リードあたりの二次(次善の)アライメント数の最大値を制限します。	--Aligner.sec-aligns	• 0〜30
sec-aligns-hard	レポートの対象となる、1リードあたりの二次(次善の)アライメント数の最大値を制限します。	--Aligner.sec-aligns-hard	• 0〜1

名前	説明	対応するコマンドライン オプション	値
sec-phred-delta	放出される二次アライメントをコントロールします。一次アライメントのPhred値範囲内にある二次アライメントのみがレポートされます。	`--Aligner.sec-phred-delta`	• 0～255
sec-score-delta	二次アライメントが許される一次アライメント未満のペアスコア閾値を決定します。	`--Aligner.sec-score-delta`	
supp-aligns	レポートの対象となる、1リードあたりの補足的（キメラ）アライメント数の最大値を制限します。	`--Aligner.supp-aligns`	• 0～30
supp-as-sec	二次フラグを付けて、補足的アライメントをレポートする必要があるかどうかを決定します。	`--Aligner.supp-as-sec`	• 0～1
supp-min-score-adj	補足的アライメントの最小アライメントスコアを増加させる量を指定します。このスコアは、ホストソフトウェアにより、「8 * DNAのマッチスコア」の式で計算されます。RNAの初期設定は0です。	`--Aligner.supp-min-score-adj`	
unclip-score	リードのエッジに到達するためのスコアボーナスを指定します。	`--Aligner.unclip-score`	• 0～127
unpaired-pen	Phred値を使用して、ペアリングされていないアライメントに対するペナルティを指定します。	`--Aligner.unpaired-pen`	• 0～255

`--enable-sampling`オプションを使ってインサート長統計の自動検出を無効にした場合、この統計を指定するには、以下のオプションをすべてオーバーライドする必要があります。詳細については、73 ページの「平均インサートサイズの検出」 を参照してください。以下のオプションは、構成ファイルの[Aligner]セクションの一部です。

オプション	説明	対応するコマンドラインオプション	値
pe-stat-mean-insert	テンプレート長の平均値を指定します。	`--pe-stat-mean-insert`	• 0～65535
pe-stat-mean-read-len	リード長の平均値を指定します。	`--pe-stat-mean-read-len`	• 0～65535
pe-stat-quartiles-insert	テンプレート長さの25パーセンタイル値、50パーセンタイル値、75パーセンタイル値を3つ、カンマで区切って指定します。	`--pe-stat-quartiles-insert`	• 0～65535
pe-stat-stddev-insert	テンプレート長の分散の標準偏差を指定します。	`--pe-stat-stddev-insert`	• 0～65535

バリアントコーラーオプション

以下のオプションは、構成ファイルの[Variant Caller]セクションで使用されます。詳細については、93 ページの 「バリアントコーラーオプション」 を参照してください。

名前	説明	対応するコマンドラインオプション	値
dn-cnv-vcf	*de novo*コールでは、CNVコールステップからのジョイント構造多型VCFをフィルターします。省略されている場合、コピー数バリアントの重複の確認がすべてスキップされます。	`--dn-cnv-vcf`	
dn-input-vcf	*de novo*コールで、*de novo*コールステップからジョイントスモールバリアントVCFをフィルターします。	`--dn-input-vcf`	

名前	説明	対応するコマンドライン オプション	値
dn-output-vcf	*de novo*コールで、フィルター後のVCFファイルを書き込むファイル位置を指定します。指定されていない場合、入力VCFが上書きされます。	`--dn-output-vcf`	
dn-sv-vcf	*de novo*コールでは、SVコールステップからのジョイント構造多型VCFファイルをフィルターします。省略されている場合、構造多型の重複の確認がすべてスキップされます。	`--dn-sv-vcf`	
enable-joint-genotyping	ジョイントジェノタイピングコーラーを有効にするには、trueに設定します。	`--enable-joint-genotyping`	• true • false
enable-multi-sample-gvcf	マルチサンプルgVCFファイルの生成を有効にします。trueに設定されている場合、DRAGENは入力として、統合gVCFファイルを要求します。	`--enable-multi-sample-gvcf`	• true • false
enable-vlrd	Virtual Long Read Detectionを有効にします。	`--enable-vlrd`	• true • false
pedigree-file	パネル間の家族関係を記述するpedigreeファイルへのパスを指定します（ジョイントコール固有）。サポートされているのは、トリオを含むpedigreeファイルだけです。	`--pedigree-file`	
qc-snp-DeNovo-quality-threshold	*de novo* SNPバリアントのカウントおよびレポートに使用される閾値を設定します。	`--qc-snp-DeNovo-quality-threshold`	

名前	説明	対応するコマンドライン オプション	値
qc-indel-DeNovo-quality-threshold	*de novo* INDELバリアントのカウントおよびレポートに使用される閾値を設定します。	`--qc-indel-DeNovo-quality-threshold`	
variant	単一のgVCFファイルへのパスを指定します。`--variant`オプションを複数機使用して、複数のgVCFフィルへのパスを指定することもできます。ファイル1つにつき1行を使用します。最大500のgVCFをサポートしています。	`--variant`	
variant-list	組み合わせる必要のあるインプットgVCFファイルのリストを含むファイルへのパスを指定します。ファイル1つにつき1行を使用します。	`--variant-list`	
vc-af-call-threshold	`--vc-enable-af-filter=true`によってAFフィルターが有効化されている場合、このオプションは、VCFでコールを放出するためのアリル頻度コール閾値を設定します。初期設定値は0.01です。	`--vc-af-call-threshold`	
vc-af-filter-threshold	`--vc-enable-af-filter=true`によってAFフィルターが有効化されている場合、このオプションは、放出されたVCFコールがフィルターされたとマークするためのアリル頻度フィルター閾値を設定します。初期設定値は0.05です。	`--vc-af-filter-threshold`	

名前	説明	対応するコマンドライン オプション	値
vc-callability-normal-threshold	体細胞コール可能領域レポートでコール可能とみなされるサイトの正常サンプルカバレッジ閾値を指定します。	`--vc-callability-normal-thresh`	
vc-callability-tumor-threshold	体細胞コール可能領域レポートでコール可能とみなされるサイトの腫瘍サンプルカバレッジ閾値を指定します。	`--vc-callability-tumor-thresh`	
vc-decoy-contigs	コンティグのカンマ区切りのリストへのパスを指定し、バリアントコール中にスキップします。	`--vc-decoy-contigs`	
vc-emit-ref-confidence	塩基対解像度のgVCF生成、またはバンド付きgVCF生成を有効にします。	`--vc-emit-ref-confidence`	• BP_RESOLUTION • GVCF
vc-enable-af-filter	体細胞モードで、アリル頻度フィルターを有効にします。初期設定値はfalseです。	`--vc-enable-af-filter`	• true • false
vc-enable-baf	bアリル頻度出力を有効にします。初期設定値はtrueです。	`--vc-enable-baf`	• true • false
vc-enable-decoy-contigs	デコイコンティグ上でのバリアントコールを有効にします。初期設定値はfalseです。	`--vc-enable-decoy-contigs`	• true • false
vc-enable-gatk-acceleration	GATKモードでのバリアントコーラーの実行を有効にします。初期設定値はfalseです。	`--vc-enable-gatk-acceleration`	• true • false

名前	説明	対応するコマンドライン オプション	値
vc-enable-liquid-tumor-mode	Tumor-in-Normal(TiN)コンタミネーションを説明するために、Tumor-normal解析の血液腫瘍モードを有効にします。初期設定値はfalseです。	`--vc-enable-liquid-tumor-mode`	• true • false
vc-enable-non-homref-normal-filter	nonhomref正常フィルターを有効にします。これは、正常サンプル遺伝型がホモ接合性リファレンスではない場合に、体細胞バリアントをフィルタリングして除外します。初期設定値はtrueです。	`--vc-enable-non-homref-normal-filter`	• true • false
vc-enable-orientation-bias-filter	方向バイアスフィルターを有効にします。初期設定値はfalseで、このオプションが無効であることを意味します。	`--vc-enable-orientation-bias-filter`	• true • false
vc-enable-phasing	可能な場合、バリアントのフェージングを有効にします。初期設定値はtrueです。	`--vc-enable-phasing`	• true • false
vc-combine-phased-variants	指定された値が0よりも大きい場合、フェージングされたバリアントをすべて、指定された値以下の距離を持つフェージングセットに結合します。初期設定値は0であり、オプションを無効にします。	`--vc-combine-phased-variants-distance`	0～65535

名前	説明	対応するコマンドライン オプション	値
vc-detect-systematic-noise	正常サンプルから系統的なノイズファイルを構築するために、高精度ランモードを有効にします。このモードは、腫瘍サンプルの解析を目的とはしていません。初期設定値はfalseです。	`--vc-detect-systematic-noise`	• true • false
vc-enable-roh	ROHコーラーとアウトプットを有効にします。初期設定値はtrueです。	`--vc-enable-roh`	• true • false
vc-enable-triallelic-filter	体細胞モードで、複対立遺伝子フィルターを有効にします。初期設定値はfalseです。	`--vc-enable-triallelic-filter`	• true • false
vc-enable-vcf-output	gVCFラン時にVCFファイル出力を有効にします。初期設定値はfalseです。	`--vc-enable-vcf-output`	• true • false
vc-forcegt-vcf	スモールバリアントコールのために強制ジェノタイピングを行います。スモールバリアントのリストを含むファイル(*.vcfまたは*.vcf.gz)が必要です。	`--vc-forcegt-vcf`	強制ジェノタイピングのためのスモールバリアントを指定する*.vcfまたは*.vcf.gzファイル。
vc-gvcf-bands	gVCFアウトプットで使用するバンドを定義します。初期設定値は、生殖細胞系列コーリングの場合`1 10 20 30 40 60 80`で、体細胞の場合`1 3 10 20 50 80`です。enable-multisample-gvcfが有効である場合、初期設定値は`5`、`20`、`60`です。	`--vc-gvcf-bands`	
vc-gvcf-homref-lod	体細胞homrefコールの検出限度を設定します。初期設定値は0.05です。	`--vc-gvcf-homref-lod`	

名前	説明	対応するコマンドライン オプション	値
vc-hard-filter	ブール式リストを使用して、バリアントコールをフィルターします。初期設定の式はDRAGENHardQUAL:all: QUAL < 10.4139;LowDepth:all: DP < 1です。	`--vc-hard-filter`	• QD • MQ • FS • MQRankSum • ReadPosRankSum • QUAL • DP • GQ
vc-max-alternate-alleles	VCFまたはgVCFにアウトプットするALTアリルの最大個数を指定します。初期設定値は1000です。	`--vc-max-alternate-alleles`	
vc-max-reads-per-active-region	ダウンサンプリングの活性領域に対するリードの最大数を指定します。初期設定値は10000です。	`--vc-max-reads-per-active-region`	
vc-max-reads-per-active-region-mito	ミトコンドリアスモールバリアントコールの活性領域に対するリードの最大数を指定します。初期設定値は40000です。	`--vc-max-reads-per-active-region-mito`	
vc-max-reads-per-raw-region	ダウンサンプリングの処理前領域1つあたりのリードの最大数を指定します。初期設定値は30000です。	`--vc-max-reads-per-raw-region`	
vc-max-reads-per-raw-region-mito	ミトコンドリアスモールバリアントコールの所定の処理前領域をカバーするリードの最大数を指定します。初期設定値は40000です。	`--vc-max-reads-per-raw-region-mito`	

名前	説明	対応するコマンドライン オプション	値
vc-min-base-qual	スモールバリアントコーラーの活性領域検出で考慮される塩基クオリティの最小値を指定します。初期設定値は10です。	`--vc-min-base-qual`	
vc-min-contig-qual	De Bruijnグラフの構築で考慮される塩基クオリティの最小値を指定します。初期設定値は10です。	`--vc-min-contig-qual`	
vc-min-tail-qual	リードのいずれかの末端で連続する塩基をトリミングするための塩基クオリティの最小値を指定します。初期設定値は10です。	`--vc-min-tail-qual`	
vc-min-call-qual	コールの放出に使用されるバリアントコールの最小値を指定します。初期設定値は3です。	`--vc-min-call-qual`	
vc-min-read-qual	スモールバリアントコールで考慮される塩基クオリティの最小値（MAPQ）を指定します。初期設定値は以下のとおりです: • 1：生殖細胞系列 • 3：体細胞T/N • 20：体細胞Tのみ	`--vc-min-read-qual`	
vc-min-reads-per-start-pos	ダウンサンプリングの開始位置ごとにリードの最小数を指定します。初期設定値は10です。	`--vc-min-reads-per-start-pos`	

名前	説明	対応するコマンドライン オプション	値
vc-min-tumor-read-qual	バリアントコーリングで考慮される主要リードクオリティの最小値(MAPQ)を指定します。	`--vc-min-tumor-read-qual`	
vc-orientation-bias-filter-artifacts	フィルタリングされるアーティファクトタイプを指定します。アーティファクト、または、アーティファクトとその逆相補は2回リストできません。	`--vc-orientation-bias-filter-artifacts`	• C/T、G/T • C/T、G/T、C/A
vc-override-tumor-pcr-params-with-normal	両方のサンプルの解析において腫瘍サンプルのパラメーターを無視し、正常サンプルのパラメーターを使用します。初期設定値はtrueです。	`--vc-override-tumor-pcr-params-with-normal`	• true • false
vc-remove-all-soft-clips	trueの場合、バリアントコーラーは、バリアントの決定に、リードのソフトクリップを使いません。初期設定値はfalseです。	`--vc-remove-all-soft-clips`	• true • false
vc-roh-blacklist-bed	指定されている場合、ROHコーラーはブロックリストBEDの領域に含まれているバリアントを無視します。	`--vc-roh-blacklist-bed`	
vc-sq-call-threshold	VCFでコールを放出するためのSQコール閾値を設定します。初期設定値は、Tumor-Normalについては3.0で、Tumor-onlyについては0.1です。	`--vc-sq-call-threshold`	

名前	説明	対応するコマンドライン オプション	値
vc-sq-filter-threshold	VCFでのフィルタリングのとおりにSQフィルター閾値マークコールを設定します。初期設定値は、Tumor-Normalについては17.5で、Tumor-onlyについては3.0です。	`--vc-sq-filter-threshold`	
vc-target-bed	指定されている場合、バリアントコーラーは、バリアントのリファレンス距離がターゲットBEDのいずれかの領域と重複しているときにのみ、そのバリアントを出力します。	`--vc-target-bed`	*.bedファイル
vc-target-bed-padding	スモールバリアントコーラーが、各ターゲットBED領域を埋めるために使用する塩基の数を指定します。初期設定値は0です。	`--vc-target-bed-padding`	
vc-target-coverage	ダウンサンプリング用のターゲットカバレッジを指定します。初期設定値は、生殖細胞系列モードでは500、体細胞モードでは50です。	`--vc-target-coverage`	
vc-target-coverage-mito	ミトコンドリアスモールバリアントコールで、開始位置が所定の位置と重複するリードの最大数を指定します。初期設定値は40000です。	`--vc-target-coverage-mito`	

名前	説明	対応するコマンドラインオプション	値
vc-tin-contam-tolerance	予想される最大Tumor-in-Normal（TiN）コンタミネーションを設定します。 0以外の値を設定した場合、血液腫瘍モードが有効になります。血液腫瘍モードを有効化した場合の初期設定値は0.15です。 血液腫瘍モードを無効化した場合の初期設定値は0です。	`--vc-tin-contam-tolerance`	
vc-excluded-regions-bed	指定した場合、BEDファイル内の領域と重複するバリアントは、ハードフィルタリングされます。	`--vc-excluded-regions-bed`	
vc-systematic-noise	AQスコア（系統的ノイズスコア）を計算するために、部位固有の系統的なノイズレベルを持つBEDファイルを指定します。	`--vc-systematic-noise`	
vc-systematic-noise-filter-threshold	系統的ノイズフィルターを適用するために、AQ閾値を設定します。初期設定値は、Tumor-Normalについては10で、Tumor-onlyについては60です。	`--vc-systematic-noise-filter-threshold`	• 0～100

CNVコーラーのオプション

以下のオプションは、CNVコーラーで使用できます。

名前	説明	対応するコマンドラインオプション	値
cnv-exclude-bed	CNV処理から除外する領域を指定します。	`--cnv-exclude-bed`	

名前	説明	対応するコマンドライン オプション	値
cnv-exclude-min-overlap	リストからターゲットを除外するために、ターゲット間隔とブロックリスト間で重複する部分の最小割合を指定します。	`--cnv-exclude-min-overlap`	
cnv-cbs-alpha	検査で変更点を受け入れる有意性レベルを指定します。初期設定値は0.01です。	`--cnv-cbs-alpha`	
cnv-cbs-eta	並べ替えメソッドを使用する際の早期停止で、連続した境界のタイプIエラー率を指定します。初期設定値は0.05です。	`--cnv-cbs-eta`	
cnv-cbs-kmax	並べ替えの小さなセグメントの最大幅を指定します。初期設定値は25です。	`--cnv-cbs-kmax`	
cnv-cbs-min-width	変更後のセグメントのマーカーの最小数を指定します。初期設定値は2です。	`--cnv-cbs-min-width`	
cnv-cbs-nmin	最大統計近似のデータの最小長を指定します。初期設定値は200です。	`--cnv-cbs-nmin`	
cnv-cbs-nperm	P値計算で使用される並べ替え数を指定します。初期設定値は10000です。	`--cnv-cbs-nperm`	
cnv-cbs-trim	分散計算でトリミングされるデータの割合を指定します。初期設定値は0.025です。	`--cnv-cbs-trim`	

名前	説明	対応するコマンドライン オプション	値
cnv-counts-method	アライメントのカウントに使用する重複メソッドを指定します。	`--cnv-counts-method`	• midpoint • start • overlap
cnv-enable-gcbias-correction	GCバイアス補正を有効にします。初期設定はtrueです。	`--cnv-enable-gcbias-correction`	• true • false
cnv-enable-gcbias-smoothing	GCビン全体でスムージングを有効にします。初期設定値はtrueです。	`--cnv-enable-gcbias-smoothing`	• true • false
cnv-enable-plots	プロットの生成を有効にします。初期設定値はfalseです。	`--cnv-enable-plots`	• true • false
cnv-enable-ref-calls	trueの場合、コピーニュートラル(REF)コールは、出力VCFファイルに組み込まれます。	`--cnv-enable-ref-calls`	• true • false
cnv-enable-self-normalization	自己ノーマライゼーションを有効にします。	`--cnv-enable-self-normalization`	• true • false
cnv-enable-tracks	表示するためにIGVにインポートできるトラックファイルの生成を有効にします。初期設定はtrueです。	`--cnv-enable-tracks`	• true • false
cnv-extreme-percentile	フィルタリングによるサンプルの除外に使用される極端な中央値パーセンタイル値を指定します。初期設定値は2.5です。	`--cnv-extreme-percentile`	
cnv-filter-bin-support-ratio	全体のイベント長に対して、ビンをサポートする範囲が指定した割合よりも小さい場合、候補イベントをフィルタリングして除外します。初期設定の割合は0.2です(20%サポート)。	`--cnv-filter-bin-support-ratio`	

名前	説明	対応するコマンドライン オプション	値
cnv-filter-copy-ratio	レポートされるイベントが出力VCFファイルでPASSとマーキングされている、約1.0を中心とする最小コピー割合閾値を指定します。初期設定値は0.2です。	`--cnv-filter-copy-ratio`	
cnv-filter-de-novo-quality	イベントのコールに使用されるPhredスケールの閾値を、発端者の*de novo*として指定します。	`--cnv-filter-de-novo-quality`	
cnv-filter-length	レポートされるイベントが出力VCFファイルでPASSとマーキングされている、塩基の最小イベント長を指定します。初期設定値は10000です。	`--cnv-filter-length`	
cnv-filter-qual	レポートされるイベントが出力VCFファイルでPASSとマーキングされている、QUAL値を指定します。	`--cnv-filter-qual`	
cnv-input	BAMの代わりにCNV入力ファイルを指定します。ファイルには、*de novo*のtarget.counts.gzまたはtn.tsv.gzが指定できます。	`--cnv-input`	
cnv-interval-width	CNV WGS処理でのサンプリング間隔の幅を指定します。	`--cnv-interval-width`	
cnv-max-percent-zero-samples	ターゲットで許可されているゼロカバレッジサンプル数を指定します。ターゲットが指定した閾値を超えている場合、ターゲットはフィルタリングで除外されます。初期設定値は5%です。	`--cnv-max-percent-zero-samples`	

名前	説明	対応するコマンドライン オプション	値
cnv-max-percent-zero-targets	サンプルで許可されているゼロカバレッジターゲット数を指定します。サンプルが指定した閾値を超えている場合、サンプルはフィルタリングで除外されます。初期設定値は5%です。	`--cnv-max-percent-zero-targets`	
cnv-merge-distance	セグメントのマージで許容されるセグメントギャップの最大値を指定します。	`--cnv-merge-distance`	
cnv-merge-threshold	2つの隣接セグメントを結合する場合のセグメントの平均差の最大値を指定します。セグメント平均値は、線形コピー割合値として表されています。	`--cnv-merge-threshold`	
cnv-min-mapq	カウントするアライメントの最小MAPQを指定します。	`--cnv-min-mapq`	
cnv-normals-file	正常サンプルのパネルで使用される単一のファイルを指定します。このオプションは何回も使用できますが、1つのファイルについては1回だけです。	`--cnv-normals-file`	
cnv-normals-list	正常サンプルのパネルとして使用するリファレンスターゲットカウントファイルのリストへのパスを含むテキストファイルを指定します。	`--cnv-normals-list`	
cnv-num-gc-bins	GCバイアス補正のビン数を指定します。各ビンは、GCコンテンツの割合を表しています。初期設定値は25です。	`--cnv-num-gc-bins`	• 10 • 20 • 25 • 50 • 100

名前	説明	対応するコマンドライン オプション	値
cnv-ploidy	正常の倍数性値を指定します。出力VCFファイルに放出されるコピー数値の推定にのみ使用されます。初期設定値は2です。	`--cnv-ploidy`	
cnv-qual-length-scale	バイアス重み付け係数を指定して、長いセグメントのQUAL推定値を調整します。初期設定値は0.9303 (2-0.1)ですが、変更する必要はありません。	`--cnv-qual-length-scale`	
cnv-qual-noise-scale	バイアス重み付け係数を指定し、サンプル変動に基づいてQUAL推定値を調整します。初期設定値は1.0ですが、変更する必要はありません。	`--cnv-qual-noise-scale`	
cnv-segmentation-mode	実行するセグメンテーションアルゴリズムを指定します。	`--cnv-segmentation-mode`	• cbs • slm • hslm • aslm
cnv-skip-contig-list	WGS解析の間隔を生成する際にスキップするコンティグ識別子のカンマ区切りのリスト。指定されていない場合、初期設定では、chrM、MT、m、chrmコンティグがスキップされます :	`--cnv-wgs-skip-contig-list`	
cnv-slm-eta	平均値プロセスが値を変更するベースライン確度を設定します。初期設定値は4e–5です。	`--cnv-slm-eta`	
cnv-slm-fw	放出されるCNVの最小データポイント数を指定します。初期設定値は0です。	`--cnv-slm-fw`	

名前	説明	対応するコマンドライン オプション	値
cnv-slm-omega	実験的分散と生物学的分散の間の相対重量を調整するスケーリングパラメーターを設定します。初期設定値は0.3です。	`--cnv-slm-omega`	
cnv-slm-stepeta	距離ノーマライゼーションパラメーターを指定します。初期設定値は10000です。HSLMのみで有効です。	`--cnv-slm-stepeta`	
cnv-target-bed	サンプルカバレッジで使用するターゲット間隔を示す、適切にフォーマットされたBEDファイルを指定します。WES解析で使用します。	`--cnv-target-bed`	
cnv-target-factor-threshold	使用できるターゲットをフィルタリングで除外するための正常サンプルのパネル中央値の下部パーセンタイルを指定します。初期設定値は、全ゲノム処理では1%で、ターゲットシーケンス処理では5%です。	`--cnv-target-factor-threshold`	
cnv-truncate-threshold	極端な外れ値を切り捨てるための閾値をパーセント単位で設定します。初期設定値は0.1%です。	`--cnv-truncate-threshold`	
cnv-use-somatic-vc-vaf	純度および倍数性の決定には、VCの体細胞SNV VAFを使用します。	`-cnv-use-somatic-vc-vaf`	

系統的なノイズBED作成オプション

通常のVCFファイルから系統的なノイズBEDファイルを作成する場合に使用できるオプションは以下のとおりです。

名前	説明	対応するコマンドライン オプション	値
vc-systematic-noise-raw-input-list	使用されるVCFのリストを生成します。VCFは1行につき1つです。	`--vc-systematic-noise-raw-input-list`	
vc-systematic-noise-germline-vaf-threshold	系統的なノイズファイル構築から可能性のある生殖細胞変異を除去するための最小バリアントアリル頻度を指定します。初期設定は指定されておらず、これはすべてのバリアントが使用されていることを示します。	`--vc-systematic-noise-germline-vaf-threshold`	• 0〜1
vc-systematic-noise-use-germline-tag	可能性のある生殖細胞変異を除去するためにDRAGEN内部生殖細胞タグ付けを使用するかどうかを決定します。`--vc-systematic-noise-germline-vaf-threshold`と相互に排他的です。初期設定値はfalseです。WGS解析については、このオプションをtrueに設定することを推奨します。	`--vc-systematic-noise-use-germline-tag`	• true • false
vc-systematic-noise-method	サンプル全体にわたって系統的なノイズレベルの計算に使用するメソッドを表します。初期設定メソッドはmeanです。	`--vc-systematic-noise-method`	• mean • max • aggregate
vc-detect-systematic-noise	正常サンプルから系統的なノイズファイルを構築するために、高精度ランモードを有効にします。このモードは、腫瘍サンプルの解析を目的とはしていません。初期設定値はfalseです。	`--vc-detect-systematic-noise`	• true • false

構造多型コーラーのオプション

名前	説明	対応するコマンドライン オプション	値
enable-sv	構造多型コーラーを有効にします。初期設定値はfalseです。	`--enable-sv`	• true • false
sv-call-regions-bed	コールする領域のセットを含むBEDファイルを指定します。オプションで、ファイルをGZIP形式またはBZIP形式で圧縮できます。	`--sv-call-regions-bed`	
sv-denovo-scoring	構造多型ジョイント二倍体コールの*de novo*クオリティスコアリングを有効にします。pedigreeファイルも提供します。	`--sv-denovo-scoring`	
sv-forcegt-vcf	強制ジェノタイピングのための構造多型のVCFを指定します。バリアントは、サンプルデータで見つからない場合でも、スコアリングされて出力VCFに組み込まれます。バリアントは、サンプルデータから直接発見された追加のバリアントとマージされます。	`--sv-forcegt-vcf`	
sv-discovery	SV発見を有効にします。`--sv-forcegt-vcf`を使用して、SV発見を無効にし、強制ジェノタイピングのインプットだけを使用すべきであることを示す場合は、falseに設定します。	`--sv-discovery`	• true • false

名前	説明	対応するコマンドラインオプション	値
sv-exome	trueに設定すると、ターゲットシーケンスインプットのバリアントコーラーが設定されます。これには、高深度フィルターの無効化も含まれます。初期設定値はfalseです。	`--sv-exome`	• true • false
sv-output-contigs	Trueに設定すると、アセンブルされたコンティグシーケンスがVCFファイルに出力されます。初期設定値はfalseです。	`--sv-output-contigs`	• true • false
sv-region	デバッグのため、解析をゲノムの特定の領域に限定します。このオプションを複数回使用して、領域のリストを作成することができます。	`--sv-region`	• 必ず、「chr:startPos-endPos」の形式で指定します。

CYP2D6コーリングオプション

名前	説明	対応するコマンドラインオプション	値
enable-cyp2d6	CYP2D6ディプロタイピングを有効にします。初期設定値はfalseです。	`--enable-cyp2d6`	• true • false

リピート伸長検出オプション

以下のオプションは、構成ファイルのRepeatGenotypingセクション、またはコマンドラインで設定できます。詳細については、176 ページの 「ExpansionHunterを用いたリピート伸長の検出」 を参照してください。

名前	説明	対応するコマンドラインオプション	値
enable	リピート伸長検出を有効にします。	`--repeat-genotype-enable`	• true • false

名前	説明	対応するコマンドラインオプション	値
specs	コールする遺伝子座を記述したリピートバリアントカタログ(specification)を含むJSONファイルへのフルパスを指定します。	`--repeat-genotype-specs`	

RNA-Seqコマンドラインオプション

名前	説明	対応するコマンドラインオプション	値
enable-rna	RNA-Seqデータの処理を有効にします。	`--enable-rna`	• true • false
annotation-file	遺伝子アノテーションファイルの供給に使用します。定量と遺伝子発現のためには必須です。	`--annotation-file,-a`	• GTF（または.gtf.gz）ファイルへのパス
enable-rna-quantification	RNA定量を有効にします。	`--enable-rna-quantification`	• true • false
rna-quantification-library-type	RNAseqライブラリーの種類を指定します。初期設定値はautodetectです。	`--rna-quantification-library-type`	• IU • ISR • ISF • U • SR • SF • A
rna-quantification-gc-bias	断片数でのGCバイアス補正を有効にします。	`--rna-quantification-gc-bias`	• true • false
enable-rna-gene-fusion	RNA遺伝子融合コールを有効にします。	`--enable-rna-gene-fusion`	• true • false
rna-gf-restrict-genes	遺伝子融合において、タンパク質コーディングまたはlncRNA以外の生物型を持つ遺伝子を無視します。	`--rna-gf-restrict-genes`	• true • false

UMIオプション

名前	説明	対応するコマンドライン オプション	値
umi-library-type	UMI修正に使用するバッチオプションを設定します。必須ではありません。	`--umi-library-type`	• random-duplex • random-simplex • nonrandom-duplex
umi-enable	UMIベースのリード処理を有効にします。	`--umi-enable`	• true • false
vc-enable-umi-solid	固形がんUMI設定を有効にします。初期設定値はfalseです。	`--vc-enable-umi-solid`	• true • false
vc-enable-umi-liquid	血液腫瘍UMI設定を有効にします。初期設定値はfalseです。	`--vc-enable-umi-liquid`	• true • false
umi-correction-scheme	UMI中のシーケンスエラーの修正に使用する方法を記述します。	`--umi-correction-scheme`	• lookup • random • none • positional
umi-correction-table	ルックアップ補正スキームのための補正テーブルへのパスを指定します。	`--umi-correction-table`	• テーブルファイルへのパス
umi-emit-multiplicity	コンセンサスリードのアウトプットの種類を設定します。	`--umi-emit-multiplicity`	• both • duplex • simplex
umi-min-supporting-reads	コンセンサスリードの生成に必要なUMIとそれにマッチする位置を持つ入力リード数を指定します。	`--umi-min-supporting-reads`	• 1以上の整数。 初期設定は2です。
umi-metrics-interval-file	ターゲットメトリックのUMIのために使用されるターゲットリージョンファイルへのパスを提供します。	`--umi-metrics-interval-file`	• 有効なBEDファイルへのパス

名前	説明	対応するコマンドライン オプション	値
umi-source	UMIのリード元である位置を指定します。	`--umi-source`	• qname • bamtag • fastq
umi-fastq	各リードに対するUMIシーケンスを持つ独立したFASTQファイルへのパスを提供します。	`--umi-fastq`	• 有効なFASTQファイルへのパス
umi-nonrandom-whitelist	ランダムではない、有効なUMIシーケンスを含むファイルへのパスを提供します。パスは、1行につき1つ入力します。	`--umi-nonrandom-whitelist`	
umi-fuzzy-window-size	指定された距離まで、UMIとアライメント位置がマッチするリードをまとめます。	`--umi-fuzzy-window-size`	• 1以上の整数。 初期設定は3です。

リソースおよび参考資料

イルミナサポートサイトのDRAGENサポートページには、追加のリソースが掲載されています。これらのリソースには、トレーニング、対応製品、およびその他の注意点などが含まれます。最新バージョンについては、必ずサポートページを確認してください。

改訂履歴

文書	日付	変更内容
文書番号:200005495 v00	2021年 7月	DRAGEN BCL変換実行要件に関する情報を追加 ALTをマスクしたハッシュテーブルに関する情報を追加 SMAコールに関する情報を更新 CYP2D6の例を更新 アライメントファイルの生成に関する情報を追加 定量およびRNA QCメトリクスを追加 体細胞CNVコールを更新 Combine gVCFのジョイント解析遺伝型オプションを削除 体細胞解析の正常サンプルのパネルを削除 Illumina Annotation Engine(Nirvana)に関する情報を更新 コンティグあたりのHet/Hom比の出力ファイルを追加 遺伝子融合検出のメトリクスおよびフィルターを更新 バイオマーカーに関する情報を更新 Poly-Gトリミングを更新 UMI入力および出力に関する情報を更新 ダウンサンプリングに関する情報を追加 構造多型の*de novo*スコアリングに関する情報を更新 体細胞モードを更新 スモールバリアントコールのミトコンドリアに関する情報を更新 メチル化パイプラインのフォーマットを更新 ORA圧縮オプションを更新 scRNA細胞ハッシングおよび機能カウントに関する情報を追加 以下のコマンドラインオプションを追加 • --ht-mask-bed • --ht-allow-mask-and-liftover • --hla-min-reads • --gg-sample-rename-mapfile • --gg-concurrency-regions • --gg-discard-ac-zero • --rna-repeat-intervals • --rna-repeat-genes • --trim-r1-5prime • --trim-r1-3prime • --trim-r2-5prime • --trim-r2-3prime • --enable-down-sampler • --down-sampler-num-threads • --down-sampler-random-seed • --down-samper-genome-size • --down-sampler-reads • --down-sampler-coverage • --vc-min-contig-qual • --vc-min-tail-qual • --vc-gvcf-bands • --vc-gvcf-homref-lod

文書	日付	変更内容
		• --enable-cram-indexing • --vc-override-tumor-pcr-params-with-normal • --sv-somatic-ins-tandup-hotspot-regions-bed • --sv-enable-somatic-ins-tandup-hotspot-regions • --ora-input2 • --ora-use-hw • --ora-print-file-info 以下のコマンドラインオプションを追加 • --cnv-segmentation-bed • --match-n-score • --methylation-mapping-implementation=multi-pass

イルミナ株式会社
東京都港区芝5-36-7
三田ベルジュビル22階

サポート専用フリーダイヤル
0800-111-5011
techsupport@illumina.com
jp.illumina.com

本製品の使用目的は研究に限定されます。診断での使用はできません。

illumina®