テクニカルノート

DRAGEN™ Bio-IT Platformを用いた集団遺伝学データの処理

大規模コホートにおけるデータ解析とバリアントコールのための推奨事項

illumina®

はじめに

全エクソームシーケンス（WES）、全ゲノムシーケンス（WGS）および下流データ処理のコストが継続的に縮小していることから、今までに例をみない規模の集団シーケンス研究が実現可能になってきています。コホートレベルのバリエーションカタログは、祖先研究、レアバリアントに関する洞察、遺伝型/表現型関連の発見、および臨床ゲノム特性のアノテーションに対する重要なリソースです。そのため、コホートのコールセットはより高い精確さが重要ですが、一方で大量のサンプルからのデータを組み合わせるにはインフォマティクスと解析に課題が残されています。

集団遺伝学データ解析

集団遺伝学（PopGen）データ処理の代表的なワークフローは、リードマッピングやバリアントコールの段階でサンプルを個別に解析することから始まり、バリアントはgVCFファイルにエクスポートされます。次に、コホート中のすべてのサンプルのgVCFファイルを集約して、遺伝型およびその信頼度メトリクスが入力された概念的マトリクスを取得します（図1）。このマトリクスは、マルチサンプルVCF（DRAGEN gVCF Genotyper）、マルチサンプルgVCF（DRAGEN/Genome Analysis Toolkit（GATK）統合gVCF）、またはデータベース（GATK GenomicsDB、GLnexus RocksDB）などの複数の形式で保存できます。いずれも、バリアントを中心としてコホート全体に対するジェノタイプコールを見渡すことを目的としています。これにより、コホート情報を使用して個々のサンプルのジェノタイプコールを改善することができます。これは、ジョイントジェノタイピングという統計学的モデルです。ただし、サンプルサイズの増加はエラーの蓄積にもつながるため、注意する必要があります。

精度についてジョイントジェノタイピングの影響に関するデータは限られています。この理由の1つには、ジョイントジェノタイピングツールをgVCF集約ツールから分離することは困難であったことがあげられます。多数のサンプルを集約することは、コホート全体で一貫した方法で異なるバリアント表現を統一する際に、特別な課題を提起します。コホートサイズが大きくなると、マルチアリルバリアントやオルタナティブアリル数も増加するため、gVCFからの全データの保存とスケーラビリティーとの間での妥協が必要になります。さらに、確立されたデータ処理ワークフローであるGATKが複雑であることも、更に困難にさせています。

DRAGENプラットフォームは、コホート解析にもシンプルなワークフローを提供し、ジョイントジェノタイピング前後の出力形式はマルチサンプルVCFファイルです（図1）。これにより、ジョイントジェノタイピングモデルの影響を直接測定できます。

本テクニカルノートでは、DRAGENプラットフォームを用いたジョイントジェノタイピングのパフォーマンスを、大規模なPopGenプロジェクトでよくある以下の3つの例で評価します。

- 35×での高カバレッジWGSサンプル
- 15×での低カバレッジWGSサンプル
- 50×での高カバレッジWESサンプル

1000ゲノムプロジェクトのフェーズ3サンプル[1]のうち、最近のリシーケンス時にGATKで生成されたコールセットに対してDRAGENプラットフォームを使用した、ベンチマーク比較を示します。コールセット精度に対する各ワークフローステージの寄与を解析し、GATK Best Practicesワークフローの一部であるいくつかのメソッドが、DRAGEN生成データに対して有益でないと考えられる理由の詳細な検証結果を示します。最後に、DRAGENプラットフォームを用いて、解析可能なバリアントを得るためのコホート処理に対する推奨事項を示します。

メソッド

インプットデータセット

WGSコホート解析は1000ゲノムプロジェクトコホートに基づいて実施されました。[2] データセットは、NovaSeq™ 6000システムを用いて、30×以上のカバレッジでシーケンスした2,504サンプルのWGSを含みます。GATKワークフローを用いた同一サンプルの処理結果は公開されており、結果を再現できます。[3,4] WESコホート解析は、CEPHコレクション（CEU）からの血族関係にない8サンプルとGenome In a Bottle（GIAB）コンソーシアムのトリオからの2サンプルを含む10サンプルからなるパネルに基づいて実施されました。[5] すべてのサンプルは、NovaSeq™ 6000システムを用いてシーケンスしました。ALT contigを含むヒトリファレンスゲノムhg38をすべての解析に使用しました。

コホート解析

WGS解析では、コホートのgVCFは、DRAGENプラットフォームv3.5.7bを用いて集約し、ジョイントジェノタイピングを実施するか、またはGATK v3.5ワークフローに従ってVariant Quality Score Recalibration（VQSR）の処理を実施しました（図1）。いずれのワークフローも染色体ごとにマルチサンプルVCFを生成します。

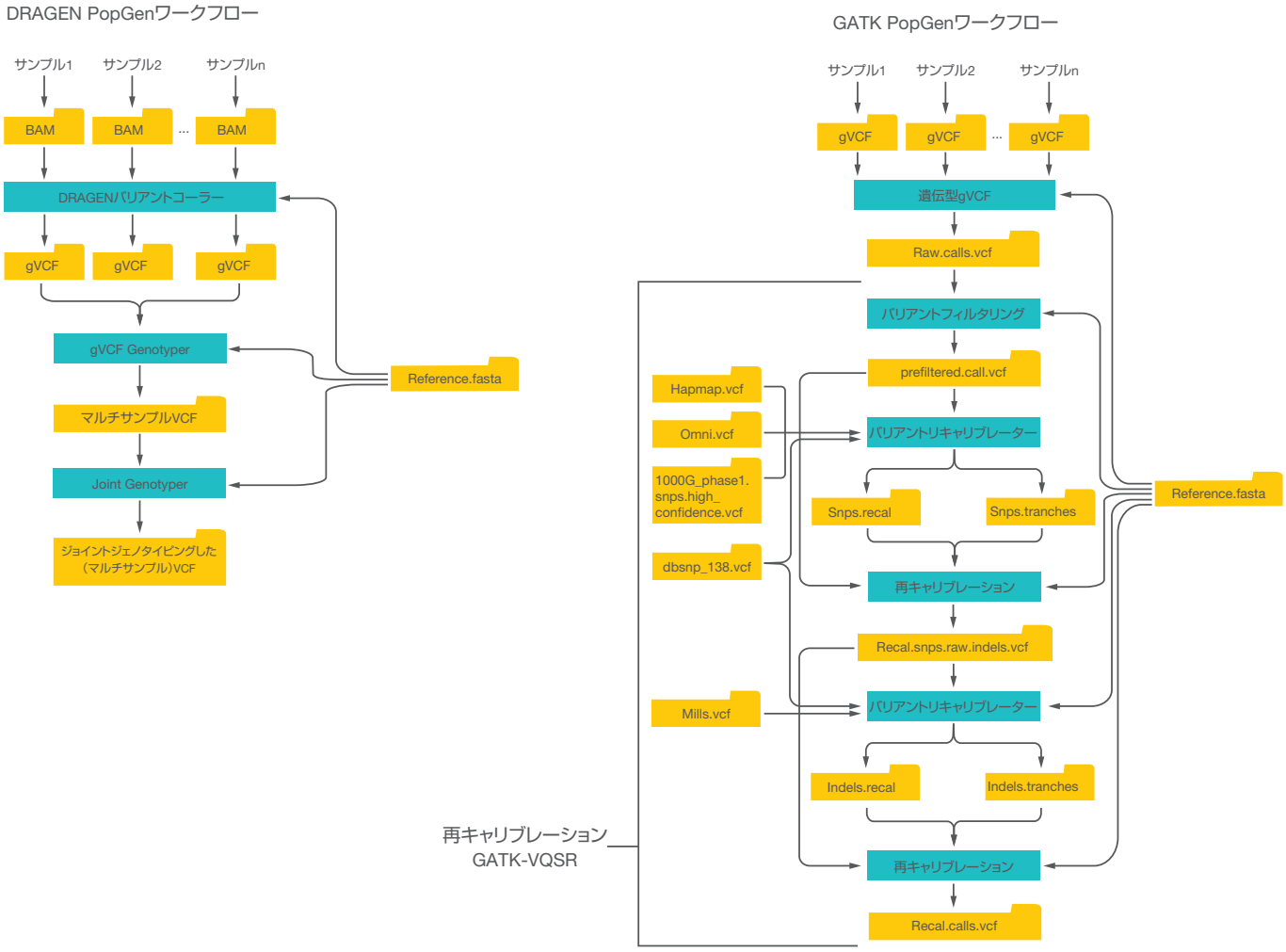


図1：DRAGENプラットフォーム（左）ワークフローおよびGATK Best Practices（右）ワークフローを用いたPopGenデータ処理と解析ワークフロー[3]：DRAGEN PopGenワークフローはgVCF Genotyper（DRAGEN-GG）を用いたコホートからのgVCFの集約、およびJoint Genotyper（DRAGEN-JG）を用いたジョイントジェノタイピングステップの2つの異なるステップからなります。DRAGENワークフローでは、再キャリブレーションステップの処理はありません。

高カバレッジWGS

高カバレッジWGSサンプルでのDRAGENプラットフォームのパフォーマンスを実証するために、DRAGENプラットフォームとGATKコールセットとの直接精度比較を実施しました。パフォーマンスは、よく特徴づけられたサンプル（NA12878）における受信者操作特性（ROC）メトリクスを用いて測定しました（NA12878はGIABによって公表された真のバリアントを持つ、オリジナルコホートの一部です）。演算コストを最小限に抑えるため、解析を17番染色体に限定しました。

ROC曲線は、さまざまな閾値での偽陽性率に対する真陽性率をプロットしました。ROC曲線下の面積は、バリアントコール精度に対するメトリクスです。

結果

マルチサンプルVCF出力からの正解サンプルNA12878を含む列を抽出し、ROC曲線をプロットすることで4つの集団データセットを評価しました。2つはGATKワークフローから、他2つはDRAGENプラットフォームからのデータセットです。

- ジョイントジェノタイピング後のパスした全バリアント（GATK-JG）*
- ジョイントジェノタイピング後のパスした全バリアントであり、かつ再キャリブレーションのみパスしたバリアント（GATK-VQSR）
- gVCF Genotyper後のパスした全バリアント（DRAGEN-GG）
- ジョイントジェノタイピング後のパスした全バリアント（DRAGEN-JG）

全体として、DRAGENプラットフォームはワークフロー構成に関わらずGATKより優れており、これはSNP（図2A）およびIndel（図2B）に対するシングルサンプルバリアントコールにおける優れた精度によるものでした。予測しなかった結果は、偽陽性の増加により、DRAGENの精度がジョイントジェノタイピング後に低下したことです（図2および図3）。現在利用できる従来のジョイントコールメソッドは、DRAGENシングルサンプルgVCFに適用すると有益にならず、不必要に高いコストとなります。これは、DRAGENプラットフォームのGenotyperがPCRで起きるエラーとパイルアップ相関エラーのモデルを組み込んでいるためです。

アプリケーションノート『Accuracy improvements in germline small variant calling with the DRAGEN platform』をご覧ください。

トリオ中のメンデルエラーは、ゲノム中の信頼度の高い領域内のバリアントに限定されないため、精度を幅広く評価する際に有用なメトリクスです。このコホート[6]の3名のうちの少なくとも1名にあるバリアント部位の合計数に関してメンデルエラーの数を評価したところ、これまでのデータと一致しました。ワークフローに関わらず、精度はDRAGENプラットフォームで向上しましたが、パフォーマンスはジョイントジェノタイピング後に低下しました（表1）。

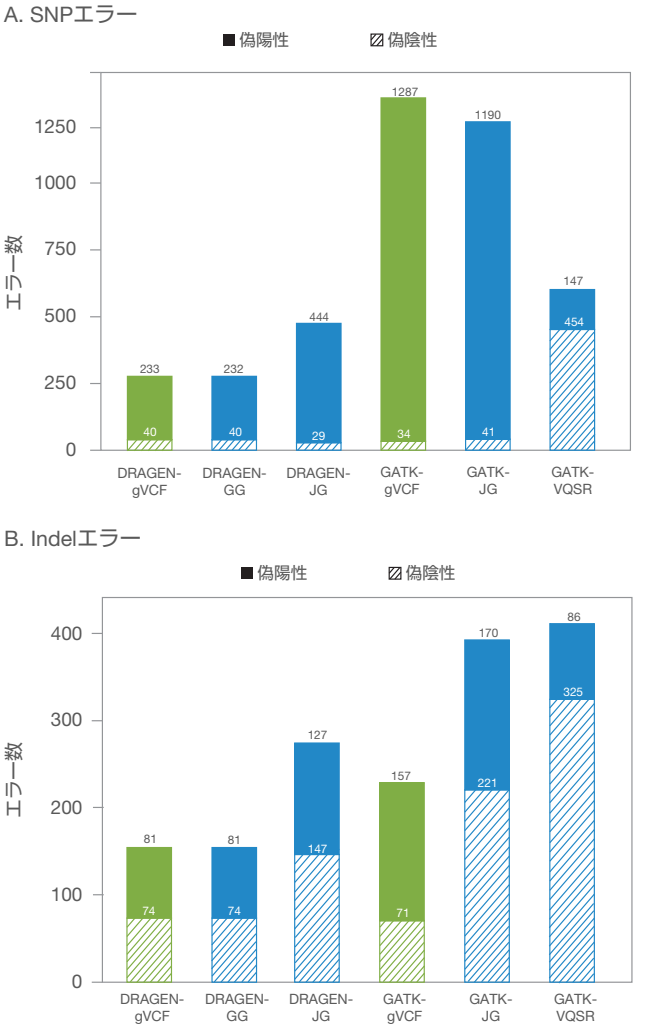


図2：高カバレッジWGSデータセットにおけるバリアントコール精度：DRAGENプラットフォーム（GG、JG）およびGATKワークフロー（JG、VQSR）でPopGen処理をした、シングルサンプルgVCF（緑）およびマルチサンプルVCF（青）における（A）SNPおよび（B）Indelのバリアントコールに対する偽陽性および偽陰性

* ジョイントジェノタイピング前のGATK集約アウトプットは使用できませんでした。

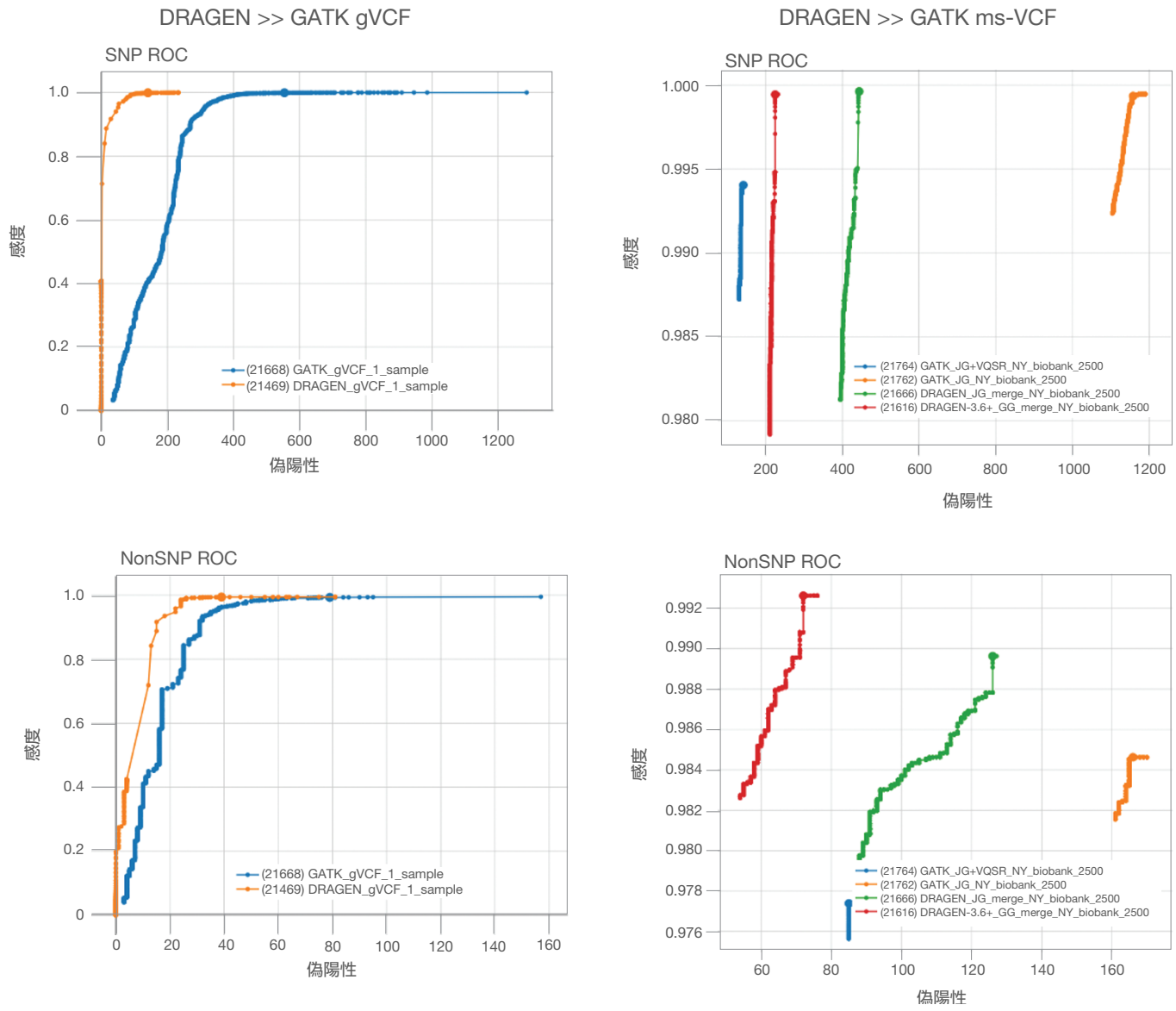


図3：高カバレッジWGSにおけるコホート解析後のROC曲線：コホート解析ワークフローからのシングルサンプルgVCF（左側）出力およびマルチサンプルVCF（右側）出力について算出したROC曲線

表1：高カバレッジWGSコホートに存在するトリオ中のメンデルエラーの算出

メンデルエラー	GATK Joint Genotyper	GATK VQSR	DRAGEN gVCF Genotyper	DRAGEN Joint Genotyper
信頼度の高い領域内	1,808/139,375 (1.30%)	833/133,195 (0.63%)	315/127,220 (0.25%)	385/127,667 (0.30%)
17番染色体全体	10,433/220,814 (4.72%)	5,272/184,275 (2.86%)	4,540/179,197 (2.53%)	5,318/186,933 (2.84%)

コホート解析におけるサンプルサイズの影響

DRAGENジョイントジェノタイピングのパフォーマンスに関するサンプルサイズの影響を、サンプル数を3、6、10、50、100に増やし、ゲノムワイドな精度メトリクスを比較することで評価しました。シングルサンプルによるベースラインメトリクスと比較した場合、SNPについては偽陰性の減少と偽陽性の増加（図4A）が、Indelについては両メトリクスの増加（図4B）が認められました。前述のように、ジョイントコールメソッドは、DRAGENシングルサンプルgVCFに対して有益ではありません。これはDRAGENのGenotyperにPCRによるエラーとパイルアップ相関エラーのモデルが組み込まれているためです。

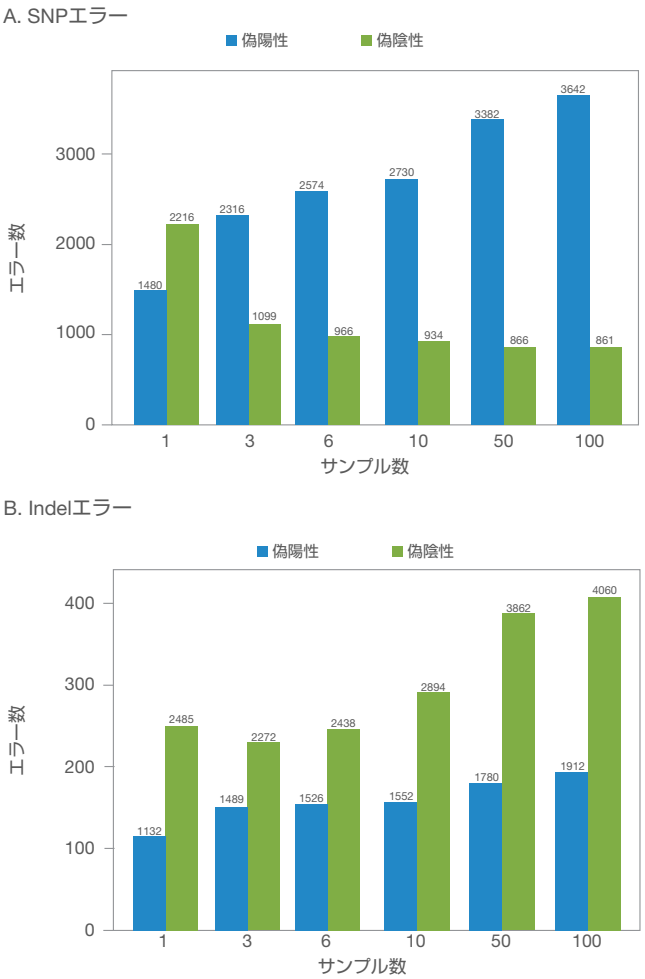


図4：ジョイントジェノタイピングに対するサンプルサイズの影響：高カバレッジWGSデータセットのサンプルサイズを増やした場合、DRAGENプラットフォームによるジョイントジェノタイピング後にプロットした（A）SNPおよび（B）Indelに対する偽陽性および偽陰性

低カバレッジWGS

低カバレッジでのジョイントジェノタイピングの潜在的な利点を検証するために、1,000ゲノムコホートからのアライメントを15×にダウンサンプリングし、これらをDRAGENプラットフォームで再処理しました。17番染色体の初めの10 Mbpからなる領域をこの解析の対象に選択しました。ダウンサンプリングしたデータのgVCFを集約し、ジョイントジェノタイピングを実施し、正解サンプルNA12878に対するROCメトリクスを用いてパフォーマンスを測定しました。

結果

低カバレッジWGSデータセットのパフォーマンスは、マルチサンプルVCFからの正解サンプルNA12878を含む列を抽出し、gVCF GenotyperおよびJoint Genotyperで処理した後にエラーカウントをプロットして測定しました。結果は、高カバレッジデータと同様であり、SNP感度の向上は特異度の低下を上回り（図5A）、Indelコールはすべてのメトリクスで悪化を示す（図5B）結果となりました。

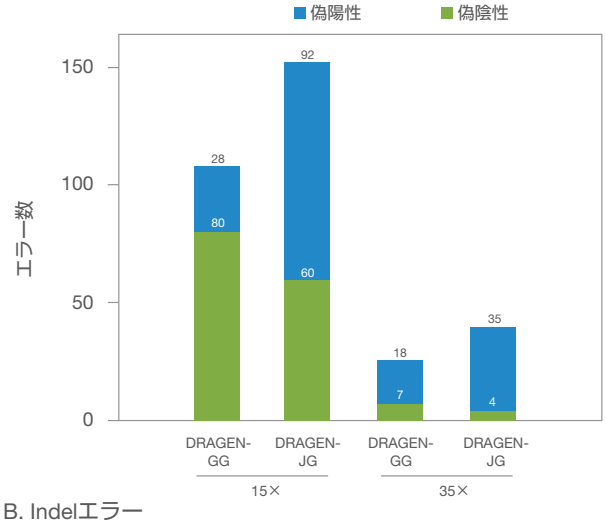
高カバレッジWESデータセット

WESデータにおけるDRAGEN Joint Genotyperのパフォーマンスは、CEU集団からの血族関係にない8名のサンプルとGIABトリオからの子ども2名のサンプルを含む10サンプルからなるパネルを用いて測定しました。ジョイントジェノタイピングは、1、3、4、6、8および10サンプルからなるサブセットで実施しました。パフォーマンスは、正解サンプルNA12878のROCメトリクスを用い、エクソームキャプチャー領域内を測定しました。

結果

異なるサブセットからのコールは、マルチサンプルVCF出力から正解サンプルNA12878を含む列を抽出し、ROC曲線をプロットして評価しました。他の解析と同様、より多くのサンプルのジョイントジェノタイピングから、明らかな利点は確認できませんでした（図6）。望ましいDRAGEN PopGenワークフローは、gVCF Genotyper実施後に終了し、ジョイントジェノタイピングステップを含めません（図7）。

A. SNPエラー



B. Indelエラー

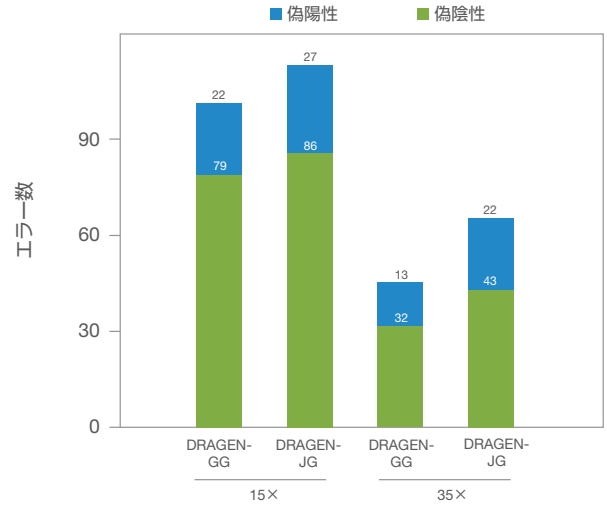
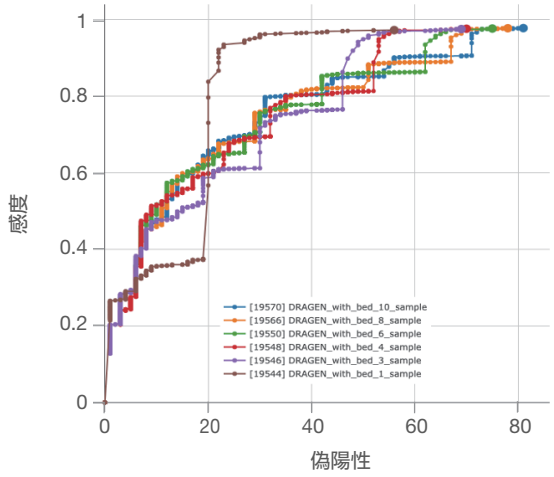


図5：低カバレッジWGSデータセットにおけるバリアントコール精度：シーケンスカバレッジ15×および35×を比較した、DRAGENプラットフォーム（GG、JG）を用いたPopGen処理後のマルチサンプルVCFにおける（A）SNPおよび（B）Indelのバリアントコールに対する偽陽性および偽陰性

A. SNP ROC



B. NonSNP ROC

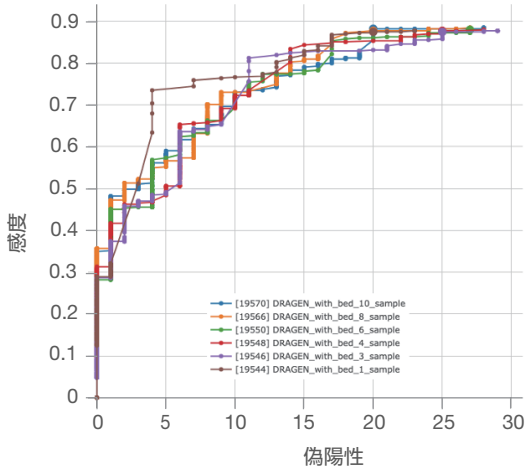


図6：高カバレッジWESにおけるDRAGEN Joint Genotyperの影響：DRAGENプラットフォームを用いたジョイントジェノタイピング後のサンプル数増加に対するROC曲線

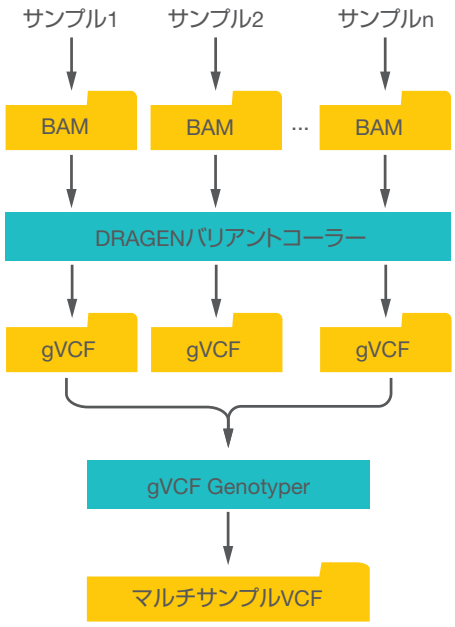


図7：推奨されるDRAGEN PopGenワークフロー

まとめ

コホートデータ処理と解析に対して確立されたGATK Best Practicesワークフローには、ジョイントジェノタイピングステップが含まれており、このステップでは、コホート情報は個々のサンプルにおけるジェノタイプコールを改善するために使用されます。しかし、本テクニカルノートに示した結果に基づくと、十分なカバレッジのあるサンプル（カバレッジ30×以上）からなる大規模コホートに対して、GATKワークフローで実施されるジョイントジェノタイピングは、DRAGENプラットフォームでの使用には推奨されません。これは、エラー発生、演算時間およびコストのリスクがあるためです。望ましいDRAGEN PopGenワークフローは、gVCF Genotyper実施後に終了し、ジョイントジェノタイピングステップを含めません（図7）。DRAGEN PopGenワークフローでは個々のgVCFの集約の後、解析準備の整ったバリアントからマルチサンプルVCFを生成します。DRAGENプラットフォームを用いたこのシンプルなワークフローにより、柔軟かつ効率的な方法で高い精度の集団コールセットが得られます。

参考文献

1. The 1000 Genomes Project Consortium; Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526:68–74. doi: 10.1038/nature15393.
2. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467:1061–73. doi: 10.1038/nature09534.
3. Intel, 2016.Infrastructure for Deploying GATK Best Practices Pipeline. intel.com/content/dam/www/public/us/en/documents/white-papers/deploying-gatk-best-practices-paper.pdf. Accessed December 01, 2020.
4. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*.2011;43:491–501.
5. Zook JM, McDaniel J, Olson ND, et al. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol*. 2019;37:561-6. doi: 10.1038/s41587-019-0074-6.
6. Roslin NM, Welli L, Paterson AD, Strug LJ. Quality control analysis of the 1000 Genomes Project Omni2.5 genotypes. *bioRxiv*. 2016;078600–078600. doi: https://doi.org/10.1101.078600.

イルミナ株式会社

〒108-0014 東京都港区芝 5-36-7 三田ベルジュビル 22 階
Tel (03) 4578-2800 Fax (03) 4578-2810
jp.illumina.com
www.facebook.com/illuminakk

販売店

Pub. No. M-GL-00561 v1.0-JPN 15MAR2022

illumina®