NovaSeq / HiSeq / NextSeq システムの多型解析結果の比較

はじめに

がんの原因遺伝子の探索や、他の遺伝性疾患の原因を探る研究において、次世代シーケンサーは、全ゲノムワイドに解析するための有用なツールとなっている。イルミナ次世代シーケンサーに使われている SBS テクノロジーは、今日に至るまで様々な改良がなされ、最新の NovaSeq システムは、1 ランで最大 60 億塩基（6Tb）にもおよぶデータ出力を可能にする。NovaSeq システムはイルミナ NGS の次の 10 年を担う大型シーケンサーであり、NextSeq システムや MiniSeq システムと同じ 2 色蛍光による塩基の検出を行う。今回、3 機種（NovaSeq、HiSeq、NextSeq システム）より得られた塩基情報を Platinum Genome と比較し、3 機種間での多型解析性能について検証を行った。

実験方法

解析には、BaseSpace Sequene Hub（BSSH）上の公開データを利用した。このうち、十分検証されている多型情報が存在する NA12878 および NA12877 を用いてシーケンスランを行ったプロジェクトを使用した（表 1）。

これらのプロジェクトは NovaSeq、HiSeq、NextSeq システムで全ゲノムシーケンスを行ったデータで、TruSeq Nano、TruSeq PCR Free ライブラリー調製キットを使用している。アライメント・バリアントコールを行った結果は、Platinum Genome の SNV や Indel などの情報と比較した。Platinum Genome はヒトゲノム全体の SNV や Indel の多型データで、多型解析（Variant call）の性能評価に使用できる高い信頼性であることが確かめられている [1]。

表 1. 比較解析に使用した BSSH で提供されている公開データ

プロジェクト名（Projects）	シーケンス条件	>Q30%
NovaSeq： TruSeq PCR-Free 350 （Replicates of NA12878）	2 × 151bp 8bp DualIndex	91.2%
NovaSeq： TruSeq Nano 550 （Replicates of NA12878）	2 × 151bp 8bp DualIndex	92.5%
HiSeq 2500 - v4 reagents： TruSeq PCR Free （4 replicates of NA12877）	2 × 126bp	90.4%
HiSeq 2000： NeoPrep Nano 550 （NA12878 - 3 replicates at 75ng plus manual at 200ng）	2 × 100bp 7bp Index	86.6%
NextSeq 500： TruSeq PCR Free WGS_ RTA2.1.3.0（NA12878）	2 × 151bp	79.4%
NextSeq 500 v2： TruSeq Nano 550（NA12878）	2 × 151bp	87.0%

解析フローを図 1 に示した。各プロジェクトに含まれるリード数、リード長は同一ではない。そのため、ゲノムでのカバレッジが 30x 程度、かつトータル塩基数が 1000 億塩基（100Gb）になるようにランダムサンプリングを行った。サンプリングには、BSSH の FASTQ Toolkit v2.2 を使用し、100Gb を含む FASTQ の作成を行った。ランダムサンプリング後の配列データは、BSSH の Whole Genome Sequencing v5.0 [2] を使用して、解析を行った。Whole Genome Sequencing v5.0 は Isaac を使用し、Human（UCSC hg19 PAR-Masked）にアライメントしたのち、多型解析には Strelka を使用した。

最後に、各プロジェクトから産出された VCF ファイルと、Platinum Genomes v2016-1.0（hg19） を BSSH の Variant Calling Assessment Tool v3.0 を用いて比較した。使用した BSSH プロジェクトに応じて、Platinum Genomes ライブラリーのうち NA12877 と NA12878 を選択して使用した。

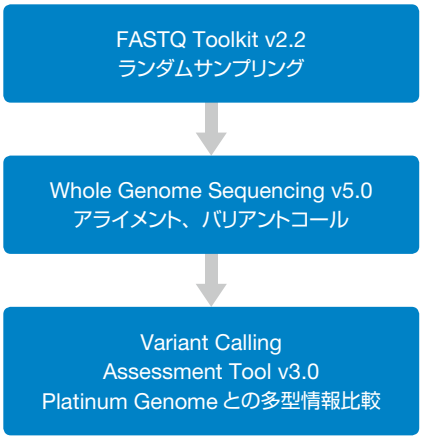

図 1：多型検出感度解析フロー

1）https://www.illumina.com/platinumgenomes.html

2）https://support.illumina.com/help/BaseSpace_App_WGS_v5_OLH_15050955_02/Content/Source/HomePages/Home_Page_WGS_App.htm

結果

Whole Genome Sequencing v5.0 を実施した際に出力されたデータから、カバレッジ、デュプリケート、常染色体カバー率など、アライメントに関する基本的なデータを表 2 にまとめた。

カバレッジは、アライメントされた配列がどれだけの厚さになるかの平均値を示している。アライメントできなかったリードが含まれる場合もあるため、それぞれの入力で若干差がある。

デュプリケートは、ペアエンドで全く同じ配列が検出される等配列の重複が見られる場合に検出される。NovaSeq システムや、HiSeq システムで v3 シーケンスキットを用いたランでは、高い傾向を示している。本資料では言及していないが、NovaSeq システムにおけるデュプリケートの割合は HiSeq X システムと同程度であるという結果が得られている。NovaSeq、HiSeq X システムに共通する整列化フローセルを使用したシーケンスケミストリーでは、HiSeq システムで v4 シーケンスキットを用いた時より高い水準でデュプリケートリードが得られる場合がある。

常染色体カバー率は、常染色体が 15 以上の厚さで読まれた部位の割合を示している。HiSeq システムで v3 シーケンスキットを用いたランにおいて 94.3% と低い値を示しているが、他のデータではおおむね 98% 程度となった。

つぎに、Variant Calling Assessment Tool v3.0 を実施して Platinum Genome に登録されている多型情報との比較を行った。SNV と Indel の精確性と、感度となるリコールについて比較した。SNV コールについては各装置間で結果に大きな差は見られなかった。Indel コールについては、SNV コールに比べて若干ばらつきが見られた。NovaSeq システムで TruSeq Nano ライブラリー調製キットを用いた場合は、Indel コールの精確性とリコールの両方で若干低い値が出る傾向が見られたが、TruSeq PCR Free ライブラリー調製キットでは他の機種と同等の結果となっている。同様の傾向は NextSeq システムで TruSeq Nano および TruSeq PCR Free ライブラリー調製キットを用いたランでも見られたが、精確性とリコールの値で傾向が違うため、単純に比較するのは難しい。

表 2. NovaSeq、HiSeq、NextSeq システムを使用した全ゲノムシーケンス結果と Platinum Genome との比較

Reagent Version WGS Library Prep Kit	NovaSeq		HiSeq		NextSeq	
	PCR Free	Nano	v4 PCR Free	v3 Nano	v1 PCR Free	v2 Nano
カバレッジ	30.8	30.7	32.4	28.9	28.8	31.0
デュプリケート	6.6%	6.6%	1.2%	7.2%	1.5%	1.6%
常染色体カバー率	98.5%	98.4%	97.3%	94.3%	96.1%	97.7%
SNV の精確性	99.9%	99.8%	99.9%	99.8%	99.5%	99.8%
SNV のリコール	97.6%	97.6%	96.9%	96.6%	96.2%	96.9%
Indel の精確性	97.4%	91.2%	97.4%	96.3%	96.3%	90.1%
Indel のリコール	95.5%	90.0%	90.5%	84.5%	88.5%	85.7%

まとめ

Platinum Genome との比較において、NovaSeq、HiSeq、NextSeq システムのいずれにおいても 99% 以上の精確性、96% 以上の感度で SNV が検出できていることがわかった。これにより、NovaSeq システムは他の装置と同等の精度で SNV コールを行うことができることがわかった。Indel コールに関しては使用するシーケンサーにより差がみられる。いずれの結果においても Indel コールの精確性が 90% 以上、リコールが 84% となっており、SNV コールには劣るものの Indel の検出も精確にできていることがわかる。Indel コールの結果の差は RTA のベースコールアルゴリズムのバージョン、使用したデータのシーケンスリード長、さらにクラスター密度などに起因すると考えられる。NovaSeq システムにて TruSeq PCR Free ライブラリー調製キットを用いると、Indel コールでの精確性が 97.4%、リコールが 95.5% の感度で検出できていることから、NovaSeq システムを使用して HiSeq、NextSeq システムと同等の結果が得られるといえる。

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

本製品の使用目的は研究に限定されます。 販売条件：jp.illumina.com/tc

Pub. No. 5019-170623-01

© 2017 Illumina, Inc. All rights reserved.
Illumina, BaseSpace, BeadArray, BeadXpress, cBot, CSPro, DASL, Design Studio, GAIIx, Genetic Energy, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, Infinium, iSelect, MiSeq, Nextera, NextSeq, NovaSeq, NuPCR, SeqMonitor, Solexa, TruSeq, TruSight, VeraCode, the pumpkin orange color, the Genetic Energy streaming bases design は、Illumina, Inc. の商標または登録商標です。
その他の会社名や商品名は、各社の商標または登録商標です。予告なしに仕様および希望販売価格を変更する場合があります。