illumina®

MiSeq™システムを用いた大腸菌シーケンス: Ion Torrent PGMシステムとの比較

はじめに

イルミナのデスクトップ型次世代シーケンサー MiSeqシステムは、可逆的ターミネーターと1塩基合成反応（SBS）を採用した実績のあるTruSeq™ ケミストリーで、高精度な塩基ごとの配列情報を産出します。これにより、繰り返し配列やホモポリマー配列も含めた全ゲノム領域にわたって、強固なベースコールをもたらします。このアプリケーションノートは、大腸菌の実験室株を用い、MiSeqシステムで得られた結果と、Ion Torrent の Personal Genome Machine（PGM）から得られた公的にダウンロード可能なデータを比較しています。クオリティ値、エラーフリーリードの割合、ホモポリマーに関連したIndel、Indelコールの予測擬陽性、GCカバレッジの正規化プロットなど、データの品質に関するポイントを中心に、ふたつのシーケンスシステムを比較しています。

結果

高品質なリードとデータ産出

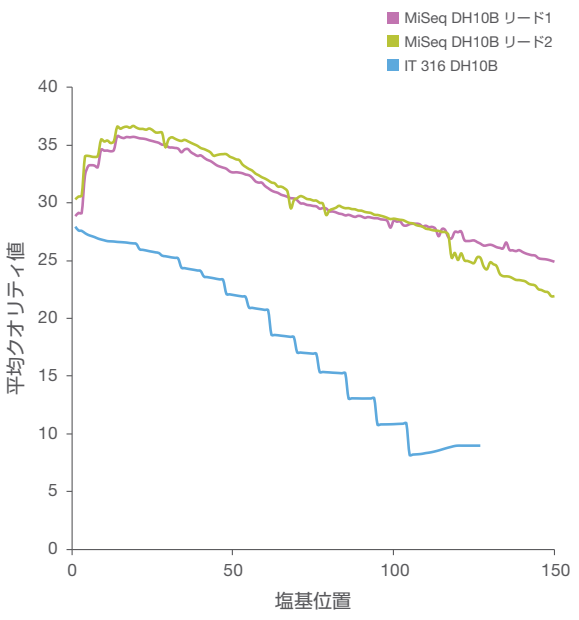
クオリティ値はある塩基が正しくコールされた確率を測定するもので、Phredに似たアルゴリズム[1]を用い、各リードの塩基ごとに測定されます。高いクオリティ値は、その塩基が正しくコールされた確率が高いものであることを意味します。図1はMiSeq 150bp x 2のランにおいて、各塩基の位置におけるクオリティ値の平均を示しています。リード1およびリード2ではQ20以上（99%以上の精度）を示しており、どちらのリードでも同様の結果となりました。一方で、Ion Torrentの316チップを用いたデータでは、クオリティ値は50bpを超えた後で大幅に下がり、PGMランで典型的に得られるリード長[2]である100bp付近にのなるとQ10（90%以下）の低いクオリティのデータが産出されています。

MiSeq 150bp x 2 のランでは、Ion Torrent PGM の 314 チップと比べて 70 ～ 80 倍以上、316 チップと比べても 10 倍以上のデータ量が得られています（表 1）。これは、MiSeq では複数のシーケンスランを行わなくても、より高い品質のデータを多く産出できることを意味します。

エラーフリーリードの割合

エラーレートはリファレンス配列とシーケンスの結果を比較することで決定されます。エラーフリーのリードの割合は、シーケンスランごとのエラーレートを示し、また使用することのできるデータ量[3]のよい指標となります。大腸菌のDH10B株を用いた結果では、MiSeqのランではエラーフリーの正確なリード数が、Ion Torrent 316チップと比べて高い数値となりました（図2）。また、別の大腸菌の実験室株であるMG1655では、MiSeqの150bp x 2および100bp x 2（150bp x 2のラン結果を100bp x 2で再解析したもの）でも同様の結果[4]となり、サンプルに用いた株およびライブラリー調製が別のものであっても、エラーレートはほとんど同じであることがわかりました。

図1：塩基位置ごとのクオリティ値

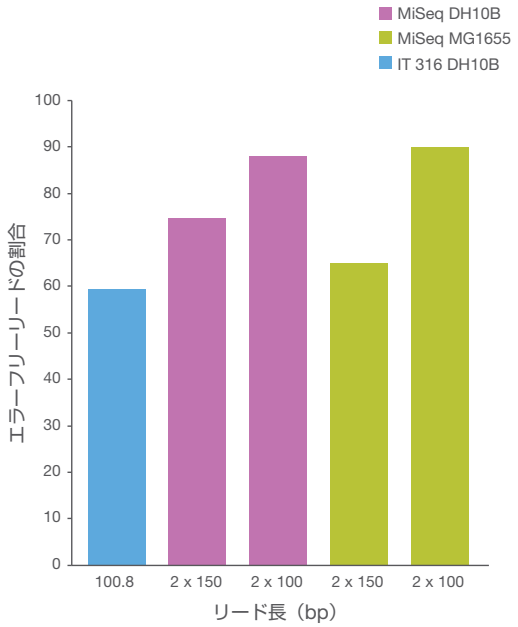


大腸菌DH10B株を用いたMiSeq 150bp x 2ラン（リード1、リード2）およびIon Torrent PGM 316チップの結果。図は塩基位置ごとの予測クオリティ値を示しています。

表1：ランのサマリー

データ	システム	ラン数	大腸菌株	ランあたりのデータ量	ランあたりの平均カバレッジ
イルミナ	MiSeq	1	K-12 DH10B	2.0 Gb	421×
イルミナ	MiSeq	1	K-12 MG1655	1.7 Gb	393×
LIFE Technologies[8]	Ion Torrent PGM（316）	1	K-12 DH10B	0.175 Gb	37×
EdgeBio[7]	Ion Torrent PGM（314）	6	K-12 DH10B	0.024 Gb	5×

図2：エラーフリーリードの割合



バクテリア株が異なっても、Ion Torrent PGM 316チップと比べてMiSeqではエラーフリーリードを高い割合で産出しています。

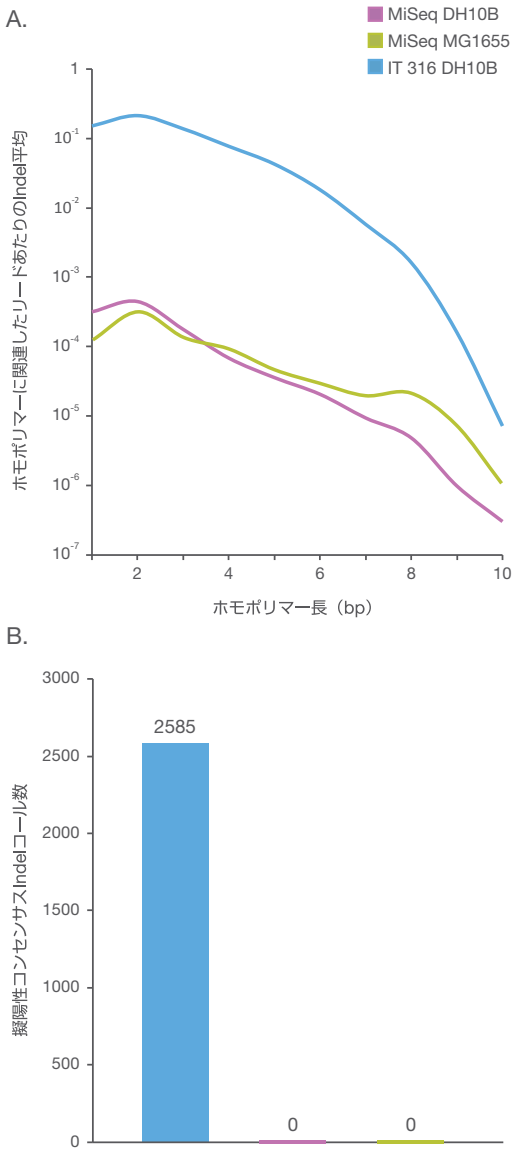
ホモポリマーに関連したIndelと擬陽性コール

ホモポリマーとは、同じ塩基が連続して並ぶ配列です。ホモポリマー領域でベースコールが正しくできないと、アライメント時に挿入（Insertion）および欠失（Deletion）の擬陽性をもたらします。6塩基以上のホモポリマーはそれほど多くはありませんが、2塩基程度の短いホモポリマーでもアライメント精度に悪影響をもたらすことがあります。フローベースの手法を用いたシーケンスとは異なり、イルミナのSBSテクノロジーは4つの可逆的ターミネーターをもつヌクレオチド（塩基）を自然競合反応させ、ホモポリマー領域でも精度の高いベースコールができるように1塩基ずつの検出を行っています。図3Aが示すように、Ion Torrent の316チップで得られた結果では、MiSeqと比べて、リードあたりのホモポリマーに関連するIndelを多く含みます。図3Bでは、全ゲノム変異解析を実施した際のIndelに関連する予測擬陽性の数を示しています。Ion Torrentの316チップではクオリティ値が20以上の予測擬陽性のIndelを多く含んだ結果となっています。MiSeqの大腸菌株DH10BおよびMG1655からの結果では、こうした擬陽性のIndelは検出されませんでした。

正規化GCカバレッジ

カバレッジの均一性はデータの包括性を測定する上で重要な指標となります。ATリッチやGCリッチのゲノムでカバレッジの均一性がないと、これらの領域の変異は十分にカバーされず、検出が困難になります。図4ではGC%ごとの正規化した平均カバレッジを示しています。データはMiSeqおよびIon Torrenの316チップを表示しています。MiSeqから得られた

図3：ホモポリマーに関連したIndelと予測擬陽性コール

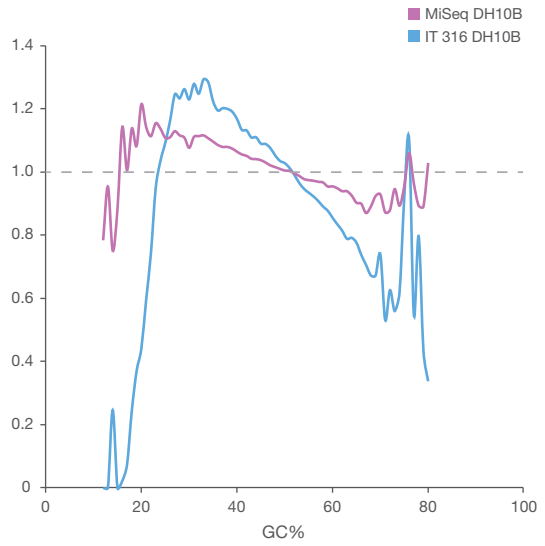


A. MiSeqからのデータはホモポリマーに関連するリードレベルでのIndelがIon Torrent 316チップと比べて低くなっています。

B. Ion Torrentのデータは予測された擬陽性コンセンサスIndelコール数が高い結果となっていますが、MiSeqでは擬陽性IndelコールはO0という結果となりました。

リードの分布では、GC%の広い範囲で、より均一な結果（正規化したカバレッジである1.0に近い）を示しています。一方で、Ion Torrentのリードでは、GC%が低い場合（0-20%）と高い場合（70-80%以上）では大きくずれが生じ、DH10Bゲノムにおいて、カバレッジの均一性が得られていないことを示しています。

図4：正規化GCカバレッジ



MiSeqとIon Torrent 316チップで解析した大腸菌DH10B株の結果を、GC%における平均正規化カバレッジプロットに示しました。理想的なGC曲線は点線に示すようにカバレッジ1.0となります。

まとめ

シーケンスの品質は、データと生物学的な関連性や、使用可能なデータ量に直接的な影響をもたらします。シーケンスの品質にはクオリティ値、リードのエラーレート、アライメント、カバレッジなど複数の要因が関与します。同じ大腸菌株を用いた結果では、Ion Torrent PGM（314, 316チップ）と比べて、MiSeqではより高い品質のデータを、より多く産出することができました。

手法

MiSeqシステムでのバクテリアゲノムシーケンス

大腸菌のK-12株であるDH10BおよびMG1655からゲノムDNAを抽出し、イルミナのTruSeqサンプル調製キットを用いてライブラリーを作成しました。ライブラリーは試薬カートリッジにアプライし、MiSeqシステムにフローセルと試薬カートリッジをセットしてランを実施しました。クラスター形成、150bp x 2ペアエンドシーケンスは、MiSeqシステム上で27時間以内に行いました。

データ解析

1次データ解析はMiSeqに搭載のコンピューター上で、特定のサーバーやコンピューターを必要とせずに行いました。Ion Torrentの結果は、公的にダウンロード可能なデータセットを使用しました（詳細はデータソースを参照）。

平均クオリティ値の比較はFASTAQファイルをPicard[5]（MeanQualityByCycle.jar）にアップロードし、平均クオリティ値を産出しました。

エラーフリーのリードは、samtools calmd[6]（samtools calmd –f reference.fa in.bam > out.sam）で産出し、out.samファイルを解析し、NMfi-rudo ga 0であるリード数を決定しました。

ホモポリマーに関連するIndelは、samtools mpileup（samtools mpileup –f reference.fa in.bam > out.txt）で産出し、リードレベルのIndelのためにout.txtファイルを解析しました。各リードレベルのIndelは、Indel両サイドの10塩基をリファレンス配列から決定し、Indelに近接する最大ホモポリマー（2以上の連続する塩基の配列）を決定しました。

予測したコンセンサスのIndelコールは下記により決定されました。

samtools mpileup -uf reference.fa in.bam | bcftools view -bvcg - > var.raw.bcf

bcftools view var.raw.bcf | vcfutils.pl varFilter > var.flt.vcf（keeping only indels w/ quality >= 20）.

GC%における平均正規化カバレッジはPicardのツールであるCollectGcBiasMetric.jarを使って計算しました。

すべてのbamファイルの入力には、マッピングクオリティ値が0以上のリードを使用し、またシステムのクオリティコントロールをパスしないリードは削除しています。

データソース

本解析は、公的にダウンロード可能なIon Torrent PGM 314および316チップを用いた、大腸菌DH10B株の結果を用いて行いました。

- EdgeBioで行われた314チップの6ランの結果[7]
- LIFE Technologiesで行われた316チップの1ランの結果[8]

参照

1. www.illumina.com/truseq/quality_101/quality_scores.ilmn
2. www.iontorrent.com/lib/images/PDFs/performance_overview_application_note_041211.pdf
3. http://www.illumina.com/truseq/quality_101/error_free_reads.ilmn
4. www.illumina.com/systems/miseq/ecoli.ilmn
5. http://picard.sourceforge.net/
6. http://samtools.sourceforge.net/mpileup.shtml
7. www.edgebio.com/data/ion/delivery/Auto_1EB-16-0039010CA/Default_Report.php
8. https://iontorrent.box.net/shared/dgjscpc8o9ic1u8uky4f

MiSeqについての詳細は
http://www.illuminakk.co.jp/product/system/miseq.shtml
をご覧ください。

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
www.illuminakk.co.jp

代理店

本製品の使用目的は研究に限定されます。

© 2013 Illumina, Inc. All rights reserved.

Illumina, illumina*Dx*, BaseSpace, BeadArray, BeadXpress, cBot, CSPro, DASL, DesignStudio, Eco, GAIIx, Genetic Energy, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, Infinium, iSelect, MiSeq, Nextera, 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。予告なしに仕様を変更する場合があります。

Pub. No. 770-2011-J012 20AUG11