OneSeq ターゲットエンリッチメント

ゲノムワイドなコピー数変化、cnLOH、挿入欠失、遺伝子変異の同時検出

アプリケーションノート

著者

Anniek De Witte

Kyeong-Soo Jeong

Arjun Vadapalli

概要

アレイ比較ゲノムハイブリダイゼーション (aCGH) プロファイリングは現在、体質性の染色体のコピー数変化を測定するゴールドスタンダードです。このテクノロジーは複数の新規微細欠失症候群のの原因となる染色体構造異常の発見に寄与してきましたが、究極の分解能である塩基対レベルでの分析はできません。1 塩基レベルの解像度をもつ全ゲノムシーケンス (WGS) を基にした初期の分析において、高解像度コピー数分析とともに SNP コールを可能にするためには、ゲノム全体にわたって深いカバレッジが必要となることが示されました。全ゲノムシーケンスで必要なシーケンシング量は、高スループット、高い費用効率を求めるほとんどの臨床研究ラボの要件には適合しません。

このアプリケーションノートでは、革新的なオールインワン SureSelect ターゲットエンリッチメントアッセイである OneSeq と、SureCall 解析ソフトウェアを紹介します。SureCall 解析ソフトウェアでは、ゲノムワイドなコピー数変化、コピー数に変化のないヘテロ接合性の消失 (cnLOH)、挿入欠失、遺伝子変異を、1 回の包括的なアッセイで検出できます。OneSeqを用い、染色体異常既知のサンプルを分析した結果、150 kb ほどの小さいサイズから染色体トリソミーまでのコピー数変化、cnLOH の領域、挿入欠失、一塩基の変異を検出できることが示されました。OneSeq ターゲットエンリッチメントは、ゲノムワイドなコピー数変化と遺伝子変異を同時に測定するための実践的ソリューションを提供します。

Agilent Technologies

はじめに

先天的な外形上の形成異常などの先天性疾患、および、知的障害、自閉症、注意欠陥・多動性障害 (ADHD) などの発達を含む精神神経疾患は、正常な発達との差異が幼児期に現れる疾患です。過去 5 年間に、これらの疾患の原因の遺伝子の同定は急速に進みました。過去の研究では主に、核型分析や、融合遺伝子を検出する蛍光 *in situ* ハイブリダイゼーション法 (FISH)、コピー数変化を検出する aCGH 法、遺伝子変異検出用のダイレクトシーケンス法と PCR 法などの手法が使用されてきました。WGS は、単一遺伝子変異から異数性までのすべての種類の異常を特定する、単一プラットフォームソリューションとなる可能性があります。しかし、深いカバレッジが必要とされる WGS にかかる費用やターンアラウンドタイムが、高スループットが求められる臨床研究ラボでの実施の妨げとなっています。ターゲットシーケンスを使用すれば、遺伝子または特定のゲノム領域のサブセットのみが配列決定され、興味あるゲノムの領域のみに時間、費用、データ保管を集中することができます。しかし、ターゲットシーケンスでは、コピー数変化をゲノムワイドに調べることはできません。OneSeq ターゲットエンリッチメントキットは、ゲノムワイドなコピー数変化、cnLOH、挿入欠失、ターゲット変異を同時に測定できるようにデザインされています。Agilent SureCall ソフトウェアのバージョン 3.0 に実装された新しいアルゴリズムにより、OneSeq データのコピー数と変異の同時解析が実現しています。

メソッド

ターゲットエンリッチメントパネルのデザイン

OneSeq ターゲットエンリッチメントキットは、Agilent SureSelect テクノロジーをベースとし、複数のプローブのセットから構成されています。第 1 セットのプローブは、実験サンプルと既知のリファレンスサンプルとの比較によってゲノムワイドなコピー数変化を検出するためのバックボーンプローブで構成されています。第 2 セットのプローブはマイナーアレル頻度の高い SNP を持つゲノム領域をターゲットとして cnLOH の検出を可能としています。第 3 セットのプローブは特定の対象領域をターゲットとすることにより、変異と挿入欠失の検出を可能としています。カタログ製品である OneSeq SS Constitutional Research Panel (図 1) は 28 Mb の合計サイズです。ゲノムワイドなバックボーンプローブによる 300 kb の解像度でのコピー数検出、疾患に関連する ClinGen 領域内での 25 ～ 50 kb の高解像度コピー数検出に対応するプローブおよび 5 Mb の解像度で cnLOH を検出するプローブ (合計 12 Mb) が含まれています。また、疾患に関連した遺伝子をターゲットとする Agilent SureSelect Focused Exome のすべてのコンテンツ (16 Mb) も含まれています。OneSeq SS Hi Res Backbone + カスタムキャプチャライブラリでは、無料のウェブベースのアプリケーションである Agilent SureDesign を使用して、ゲノムワイドの CNV バックボーンに最大 12 Mb まで任意のカスタムターゲット遺伝子パネルを追加できます。

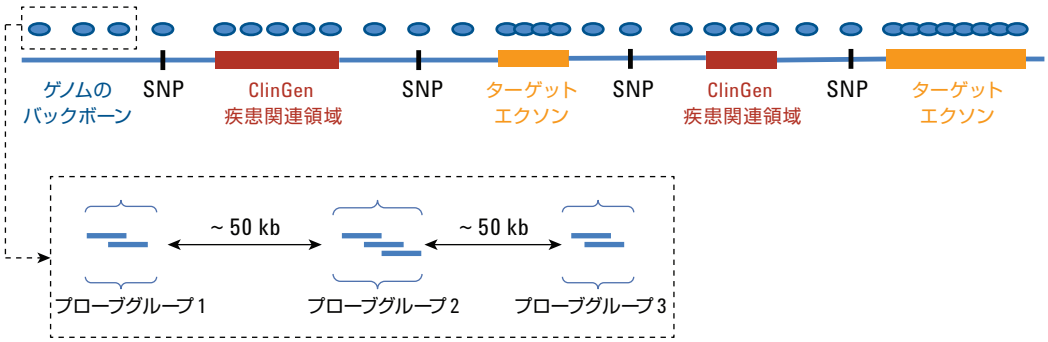

図 1. OneSeq ターゲットエンリッチメント用に使用するプローブデザインの概要。

サンプル前処理

Agilent SureSelectXT ターゲットエンリッチメントシステム · イルミナペアエンドマルチプレックス対応プロトコル Version B.1 に従い、Coriell Cell Repository (http://www.coriell.org/) 社から入手した 6 つの DNA サンプル、NA03997、NA11419、NA08254、NA04592、NA02948、NA20409 を用いて実験をおこないました。200 ng の DNA スタート量のプロトコルを使用し、キャプチャ後の PCR は 10 サイクルとしました。アジレントの Male および Female のリファレンス DNA を平行して実験し、Reference としてデータ解析で使用しました。各キャプチャ済みライブラリを Illumina HiSeq 2500 2 × 100 bp プラットフォームにロードしました。1 サンプルにつき 7 Gb のシーケンスで適切な深度とカバレッジが得られました。

データ解析

生のイメージファイルを Illumina ベースコーリングソフトウェアによりデフォルトパラメータで処理し、FASTQ ファイルを作成しました。FASTQ ファイルを Agilent SureCall ソフトウェア v3.0 にインポートしました (図 2)。アダプタシーケンスを除去した後、SureCall に組み込まれた BWA アライメントアルゴリズムを使用してリードをゲノムに対してアライメントしました。

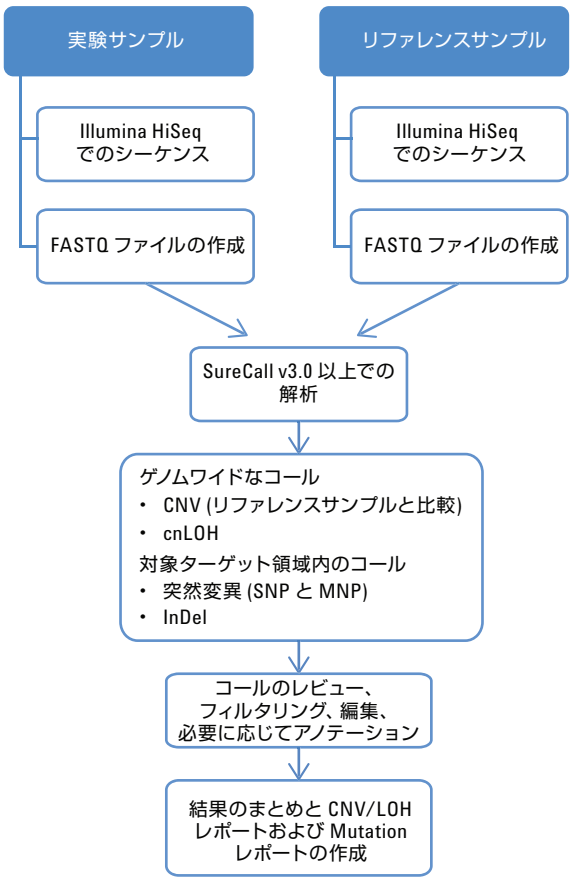

図 2. OneSeq 解析を Agilent SureCall で実行するステップ。

コピー数変化は、実験サンプルと既知のリファレンスサンプルを比較することによって検出されます (図 3)。はじめに、異常値を原因とするノイズを最小に抑えるために、プローブによってカバーされるゲノム領域にわたってリード分布の central tendency を算出するための summarization method を適用します。コピー数変化は、プローブでカバーされる各領域に対して、サンプルの Read Depth をリファレンスの Read Depth で割って計算されたログ比として示されます。ログ比を Undecimated wavelet transform 処理することによって、急激な変化がおきている位置またはブレークポイントを検出します。変換されたログ比を、さまざまなゲノム長スケールで分析します。検出されたブレークポイントを組み合わせランク付けした後、False Discovery Rate を使用して統計学的有意性の一定のしきい値をパスするブレークポイントのみ選択します。有意とされた変化領域は、増幅または欠失の候補領域として、さらに細かい解像度で調べます。

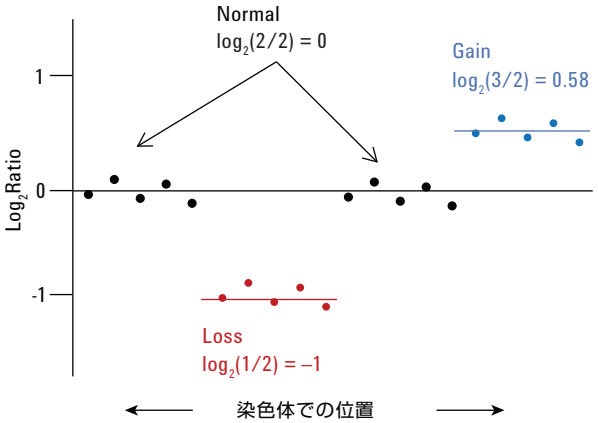

図 3. Agilent SureCall ソフトウェアでのコピー数変化の測定。サンプルシーケンスの Read Depth と対照シーケンスの Read Depth の $\log_2$ 比が染色体での位置に沿ってプロットされています。コピー数変化がない場合、$\log_2$ 比は 0 (黒の点) に、1 コピー欠失の場合、$\log_2$ 比は –1 (赤の点) に、1 コピー増幅の場合の $\log_2$ 比は 0.58 (青の点) に対応します。

点突然変異や挿入欠失のコールにはアジレント社内で開発した SNP コーリングアルゴリズム SNPPET を使用しました。SNPPET アルゴリズムには 2 段階の基本ステップがあります。最初のステップは変異のクイックサーチで、2 つのモデルで遺伝子座が評価されます。1 つめのモデルは検討中の塩基について、すべての non-reference allele は、シーケンスエラーが原因であると仮定し、2 つめのモデルはそれぞれの non-reference allele が真の変異であるとみなします。次のステップは、変異の近傍での詳細なローカルサーチです。可能性のあるすべての変異の組み合わせをハプロタイプとして評価し、近傍の変異サイトとあわせて調整します。

OneSeq バックボーン内でカバーされているマイナーアレル頻度が高い SNP は、cnLOH を決定するために使用されます。ヘテロ接合 SNP コールが統計学的に有意に少ないゲノム領域を特定することによって、cnLOH または UPD の領域を決定します。LOH アルゴリズムは、始めに、利用可能な SNP 位置で測定されたアレル頻度を使用して、サンプルを既知の母集団に 99 % の信頼度で割り当てます。サンプルを既知の母集団に割り当てることができない場合、代わりに UCSC から利用できる平均ヘテロ接合度の割合を使用します。次に、sequential Fisher's test を使用して、ヘテロ接合性が欠失した SNP が濃縮されているゲノム領域のスコアを付けます。最終的な LOH スコアは、候補となる SNP の位置に存在するかもしれない挿入欠失および複数のアレルの存在を考慮しています。アルゴリズムの詳細な説明は、SureCall ヘルプシステムで確認できます。

結果と考察

QC データ

8 つのすべてのサンプルについて、ターゲット ± 100 bp のリードの割合は 75 % 以上で、Duplicate リード数は非常に低くなりました (図 4)。これは、Human All Exon V5 キットなど他の SureSelect ターゲットエンリッチメントキットの結果と類似しています。ターゲットの塩基の 95 % 以上が最小 20 のリードでカバーされる高いカバレッジが得られました。最小 50 または 100 リードでカバーされる塩基の割合は想定通りに低くなりました。シーケンス量を 7 Gb から 10 Gb に増やすと、50 リード以上でカバーされる塩基の割合は 90 % 以上になりました (データは示していません)。

大きな CNV の検出

既知のコピー数異常がある複数のサンプル内で、期待される染色体全体のコピー数変化を検出することができました。図 5 は、Coriell 製サンプル NA02948 中の核型 47、XY、+13 のトリソミー 13 の検出を示しています。染色体全体で測定された平均 $\log_2$ 比は、シングルコピー増幅として期待される $\log_2$ 比の 0.58 に近い値になりました。

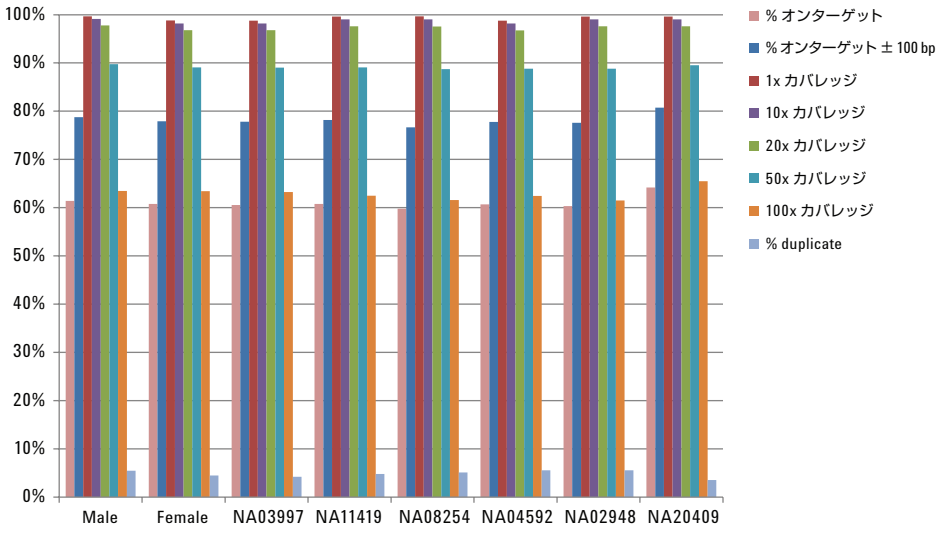

図 4. OneSeq SS Constitutional Research Panel のオンターゲット % とカバレッジ。

図 5. Agilent SureCall ソフトウェアでの OneSeq コピー数データ解析が示す、Coriell 製サンプル NA02948 (47,XY,+13) 中の 13 番染色体のトリソミー。赤の十字のプロットがそれぞれ生データの点を表しています。ChinGen の疾患関連領域内はデータポイント密度が高いことがわかります。青の陰影と線はコピー数変異のコールで、染色体 13 番の 1 コピー増幅を示しています。

aCGH との比較

OneSeq により得られたコピー数プロファイルと aCGH で得られたコピー数プロファイルを比較しました。CGH データは Agilent CGH+SNP 4 × 180K マイクロアレイを用いました。表 1 は、2 つのメソッドにより Coriell 製サンプル NA08254 で取得された、150 kb 以上の CNV コールの比較を示しています。両方のメソッドで、同じ 8 つの CNV を検出できました。Aberration の始まりと終わりのゲノム位置は、aCGH プローブと OneSeq プローブの配置が異なるため一致せず、Aberration のサイズもわずかに異なりました。図 6 と 7 は、13 番染色体での 13 Mb の欠失と 6 番染色体での 370 kb の欠失の比較データを示しています。

表 1. aCGH と OneSeq により Coriell 製サンプル NA08254 で検出された 150 kb よりも大きい CNV。

染色体	Aberration タイプ	aCGH Aberration サイズ (kb)	OneSeq Aberration サイズ (kb)	OneSeq 平均 log_2 比
chr13	欠失	12427	13335	-0.89
chr15	欠失	2240	1667	-0.37
chr16	欠失	772	863	-0.43
chr14	増幅	987	544	0.54
chr6	増幅	370	372	0.61
chr2	欠失	828	307	-0.46
chr17	増幅	163	201	0.49
chr22	増幅	172	191	3.00

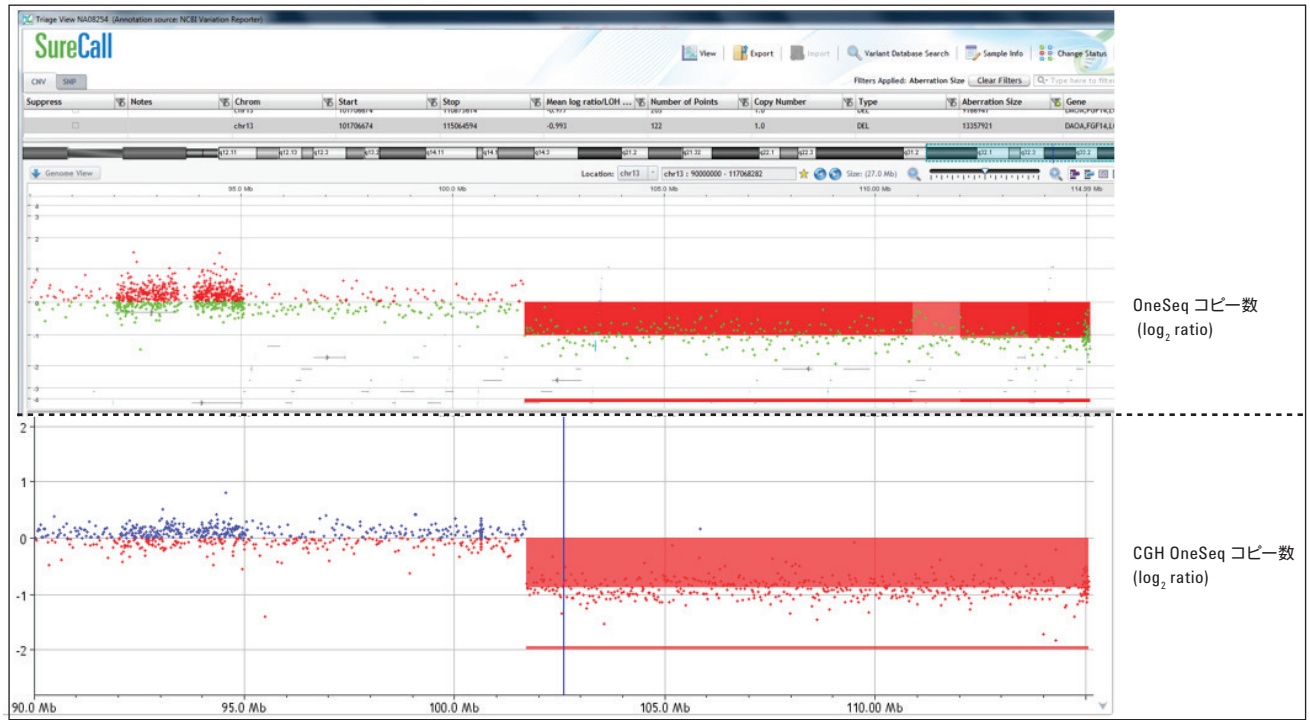

図 6. Agilent SureCall ソフトウェアでの OneSeq コピー数データ解析 (上のパネル) と Agilent CytoGenomics ソフトウェア v3.0 での aCGH コピー数データ解析 (下のパネル) は、Coriell 製のサンプル NA08254 中の 13 番染色体の 13 Mb 欠失を示しています。

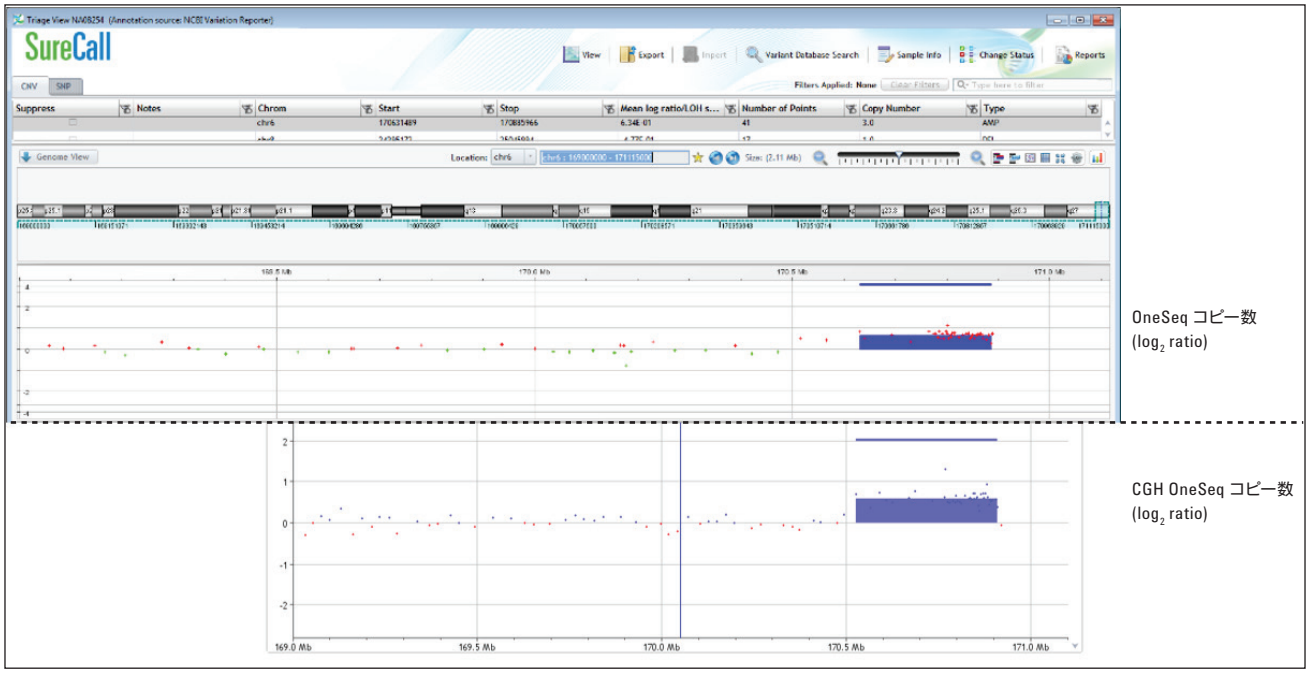

図 7. Agilent SureCall ソフトウェアでの OneSeq コピー数データ解析 (上のパネル) と Agilent CytoGenomics ソフトウェア v3.0 での aCGH コピー数データ解析 (下のパネル) は、Coriell 製のサンプル NA08254 中の 6 番染色体の 370 kb 増幅を示しています。

cnLOH の検出

図 8 は、Coriell 製のサンプル NA20409 中の UPD 15 (片親性ダイソミー) の検出を示しています。ほとんどすべての SNP の B-allele frequency は 0 % または 100 % で実質的にはヘテロ接合 SNP は染色体全体で存在しなかったため、この UPD コールは高い信頼度で検出されました。

図 8. Agilent SureCall ソフトウェアデータ解析でのコピー数および LOH データ解析による Coriell 製サンプル NA20409 中の UPD 15 を示しています。上のパネルはコピー数データです。赤および緑のそれぞれの十字は生データの点を表しています。セントロメアに近い既知の一般的な CNV を除いて染色体全体が二倍体です。下のパネルは LOH データを示しています。各水色の点は SNP の B-allele frequency (BAF) を示しています。水色の陰影は 15 番染色体の UPD を示しています。

変異と挿入欠失の検出

高い Read Depthにより、すべての Coriell 製サンプル中の対象ターゲット領域内の一塩基の変異と挿入欠失の検出が可能であることが示されました。図 9 は、メチル化 CpG 結合タンパク質 2 (MECP2) をコードする遺伝子のヌクレオチド 1160 で始まる 26 塩基対の欠失を保有することがわかっている、Coriell 製サンプル NA16382 の 26 bp 挿入欠失の検出例を示しています。

結論

OneSeq target エンリッチメントと Agilent SureCall ソフトウェアの組み合わせは、ソフトウェアで設定された各ウィンドウ内のシーケンスリードの数をサンプルとコントロールリファレンスの間で比較することによって、コピー数変化を高解像度で検出する一貫したメソッドを実現しています。OneSeq は、コピー数変化の解析だけでなく、cnLOH、SNP、InDel の検出にも対応します。WGS とは異なり、OneSeq では大量のシーケンスを必要としません。OneSeq は、費用対効果、スループットを保持しながら現行の複数のテクノロジーを統合したもので、単一遺伝子変異から異数性までさまざまな異常の臨床研究のためのシングルプラットフォームソリューションです。

図 9. Agilent SureCall ソフトウェアでの変異と挿入欠失の同定。Coriell 製サンプル NA16382 の MECP2 遺伝子中の 26 bp の挿入欠失 (右側) とヘテロ接合 SNP (左側) を示しています。

ホームページ:
AgilentGenomics.jp

カストマコンタクトセンタ
フリーダイヤル 0120-477-111

アジレント・テクノロジー株式会社

Published in Japan, March 11, 2015
5991-5631JAJP

Agilent Technologies