Twist ターゲットエンリッチメントパネルを用いたキャプチャベースの SNP ジェノタイピング

要約

一塩基多型 (SNP) を用いた大規模なジェノタイピング (遺伝子型判定) の主な手法として、長い間アレイが用いられてきました。本アプリケーションノートでは、次世代シーケンス (NGS) により数十万個のマーカーを識別するために Twist カスタムターゲットエンリッチメントパネルをどのように設計できるのかをご紹介します。バリアントコール (バリアントの検出) の性能はゲノムジェノタイピングの基準を用いて評価しました。アレイと直接比較すると、ターゲットエンリッチメントパネルはバイアスを最小限に抑えて正確にジェノタイピングできることを示します。SNP、およびインデル (挿入欠失) のジェノタイピングは、今や全エクソーム解析と同じプラットフォームで実施することが可能であり、費用、時間、労力を低減させることにつながります。

はじめに

過去 20 年間、ジェノタイピングアレイは、一塩基多型 (single nucleotide polymorphism、SNP) の大規模解析や個人の遺伝子構造解析に役立てられてきました。この基盤技術によって、進化ゲノミクスや遺伝性の複雑疾患から個別の遺伝子解析や個別化医療に至るまで、さまざまな分野において理解が進んできました。近年、次世代シーケンス (NGS) はコストが低下したことで、ジェノタイピングにおける魅力的な選択肢となっています。NGS はバリアント周辺の完全な配列情報を提供できるため、ジェノタイピングの対象が SNP のみに留まらず、複対立遺伝子領域、挿入、欠失、その他の構造多型にまで拡張されます。また、あるサンプル中に存在する可能性のある特定のバリアントや遺伝子型のためにプローブを作製する必要がないため、NGS はアレイの固定されたテンプレート形式よりも高い柔軟性をもたらします。

しかし、ターゲットシーケンスは規模が大きい場合の性能に関して問題があるため、いまだ完全にはマイクロアレイにとって代わるものとはなっていません。このため、バリアントの情報を得るために、エクソームシーケンスとアレイベースのジェノタイピングがしばしば、同一サンプルに対してそれぞれ個別のワークフローとして独立して実施されます。

本アプリケーションノートでは、シーケンスによるジェノタイピングに用いるための、Twist カスタムパネル設計アルゴリズムを活用した約 240,000 個に及ぶ SNP 用ターゲットリッチメントパネルの作製についてご紹介します。Twist カスタムパネルは、さまざまなパネルサイズ、ターゲット領域、マルチプレックス要件を網羅して設計・構築することができ、いずれも一貫して卓越した性能を発揮します。ターゲットエンリッチメントパネルは、プローブシーケンスとのミスマッチにキャプチャ効率の低下が小さい許容性があることがすでに示されています (**図 1**)。対応するアレイベースのジェノタイピング結果に対して、ゲノムジェノタイピング基準を用いてパネルの性能を評価し、バリアントコール (変異の検出) の精度および感度が 99% を超えることを示します。また、GC 含量、参照アレル (リファレンス対立遺伝子) バイアス、さまざまな集団への適用性などのバイアスを慎重に評価し、バイアスを最小限に抑えた正確なジェノタイピングであることを示します。すなわち、ジェノタイピングとエクソーム解析を統合して一元化した、それぞれ個別に実行するよりも大幅な費用削減につながるワークフローについてご紹介します。

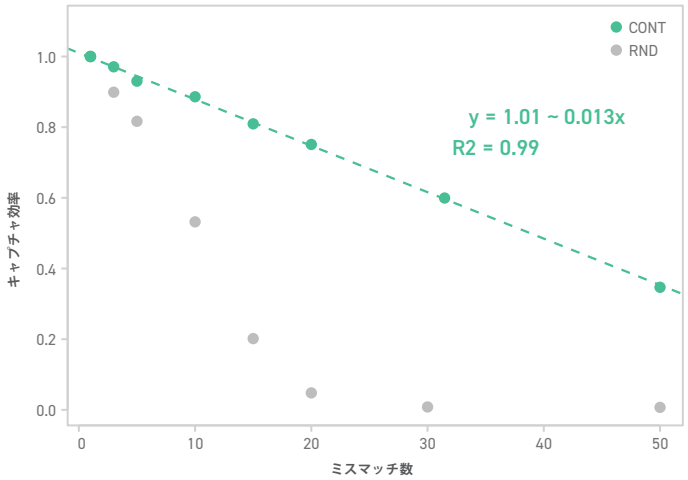


図 1：ランダムに分離 (RND) および隣接 (CONT) したミスマッチに対する Twist ターゲットエンリッチメントパネルのキャプチャ効率の堅牢性 (完全一致キャプチャに対する相対値)。詳細は、ハイブリダイゼーションによる DNA キャプチャにおけるミスマッチの影響を調べた当社のホワイトペーパーに示す。

方法

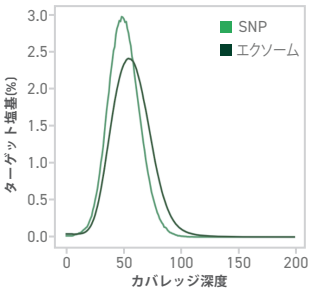
ジェノタイピングパネルの設計

シーケンスによるジェノタイピング (genotyping by sequencing; GBS) に対する Twist カスタムターゲットエンリッチメントパネルの適用性を評価するために、一般的なジェノタイピングアレイ (Illumina Infinium Global Sequencing Array、GSAv2) に含まれるバリアントに対するコンセプト実証用 SNP パネルを設計しました。エクソームとともに実施したときの GBS の性能測定を可能にするため、ミトコンドリア SNP と 250 bp 未満のバリアントを遺伝子から除去した後、約 240,000 個の SNP が残存しました。これは、Genome in a Bottle (GiAB) コンソーシアムが開発した高品質領域を参照すると、ショートリードのシーケンスに適していると判断されます。

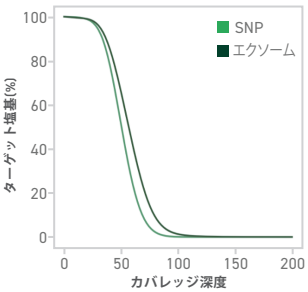
ジェノタイピングの性能評価

キャプチャ実験は、SNP パネルを個別に、または Twist Human Core Exome パネルに対するスパイクインとして用いて、Twist 標準ハイブリダイゼーションプロトコルに基づいて実施しました。いずれの実験も、Coriell 社から入手した欧州大陸、アジア大陸、アシュケナージ系を網羅するゲノム DNA サンプルを用い、反復して実施しま

カバレッジ分布



累積カバレッジ



キャプチャ指標

平均ターゲットカバレッジ	51
fold 80 ベースペナルティ	1.29
オフプローブ率	33%
20x カバレッジ（%）	99%
30x カバレッジ（%）	95%
ゼロカバレッジ	0.005%
Duplicate 率	2.5%
AT ドロップアウト	0.39
GC ドロップアウト	0.01 未満

図 2：150x の raw シーケンスカバレッジ後のターゲット SNP ジェノタイピングに対するキャプチャ性能。グラフはターゲット塩基に対するカバレッジ分布を示し、SNP およびコアエクソームの両パネルを比較した。カバレッジ分布：カバレッジが所定のレベルにある塩基の割合。累積カバレッジ：カバレッジが所定のレベル以上である塩基の割合。表は、Picard を用いて得られたその他のキャプチャ指標を示す。

SNP ジェノタイピング（TE SNP パネル）

サンプル	反復	精度	感度
NA12878	1	99.90	99.79
NA12878	2	99.91	99.81

SNP ジェノタイピング（アレイベース）

サンプル	精度	感度
NA12878_1	99.59	99.35
NA12878_2	99.59	99.30
NA12878_3	99.58	99.37
NA12878_4	99.58	99.03

インデルジェノタイピング（TE SNP パネル）

サンプル	反復	精度	感度
NA12878	反復 -1	91.07	90.39
NA12878	反復 -2	90.47	90.32

表 1：単一サンプル（NA12878）のアレイベースのジェノタイピングと比較した、Twist SNP パネルのジェノタイピング性能

した。それらは細胞株 NA12878、NA24694、NA24143 で構成されており、GiAB により包括的に評価され、米国国立標準技術研究所によってジェノタイピング用基準とされているものです。

シーケンスは、2 x 75 bp のリード長を有する NextSeq500/550 High Output キットを用いて、Illumina NextSeq プラットフォームで実施しました。マッピングクオリティ 20 以上で BWA（Li および Durbin、2009）を用いて、ヒトゲノムへのアライメントを（hg19 アセンブリに基づき）行いました（元の GSAv2 アレイは hg19 アセンブリに対して設計されています）。バリアントコールは、GATK v3.5 を用いたベストプラクティスのワークフロー（Van der Auwer ら、2013）にて実施しました。また、第三者プロバイダにより GSAv2 アレイおよび Genome Studio 2.0 を用いて、GBS に用いられる各同一サンプルの等量分割試料に対するアレイベースのジェノタイピングが反復実施され、ジェノタイプコールを生成しました。さらに、ツール（Rayner、McCarthy、ASHG、2011）を用いて、Illumina の top／bottom 表記からプラス鎖／マイナス鎖への変換を行いました。ゲノミクスと健康のための世界連合（Global Alliance for Genomics and Health; GA4GH、Kruche ら、2019）によって確立された基準パイプラインおよび推奨を用いて、対応するターゲットに対してのシーケンスまたはアレイをベースとしたジェノタイピングを、絶対的な基準として GiAB により公開されている高信頼性のコールに対してそれぞれ比較しました。

参照アレルのバイアスの評価

ヘテロ接合型 SNP 部位について変異アレルを含むリードの割合を計算し、総リード数が等しい部位において等しい確率を持つアレルのサンプリング期待値と比較しました（二項分布。p = 0.5、n = 各 SNP 遺伝子座でマッピングするリード数）。

結果

キャプチャ性能およびターゲットエンリッチメントの指標

Twist Human Core Exome Panel と同様の方法で SNP（24 万）パネルを設計し、まず各パネルのプローブ塩基数に対して 150 回の raw シーケンスで SNP ターゲットを独立にキャプチャすることにより、Twist エクソームパネルと比較しました。

SNP パネルは、優れたキャプチャ均一性（fold 80 ベースペナルティ 1.35。これに対しエクソームパネルでは 1.32）および低い Duplicate 率（2.5% 対 2.9%）を示し、Twist Custom Panel デザインに期待される極めて高品質な基準に合致しました。さらに、AT および GC のドロップアウトなどターゲット特有の指標はエクソームパネルの値と同等であり、ゼロカバレッジの SNP 部位はわずか 0.005% でした（**図 2**）。

オフターゲットキャプチャの増加（SNP では 33%、エクソームでは 15%）が観察され、それはカバレッジ分布のピークにわずかなシフトをもたらしています（**図 2**）。それにもかかわらず、プローブでカバーされるすべての塩基ではなく、SNP 部位のみをターゲットとして着目すると、SNP ターゲットの指標はエクソームの指標と一致し、20x カバレッジおよび 30x カバレッジでそれぞれ 99%、95% となり、キャプチャ均一性は若干増加しました（1.29 の fold 80 ベースペナルティ、狭いカバレッジ分布）。

高感度かつ正確なジェノタイピング

パネルのキャプチャ性能の検証（**図 2**）に続いて、絶対的な基準として調べられた 3 種類のサンプルそれぞれにおける GiAB コールを用いて、同じ SNP に対するアレイベースのコール率と比較することで GBS 指標を評価しました。GBS における 150x のシーケンスに基づく

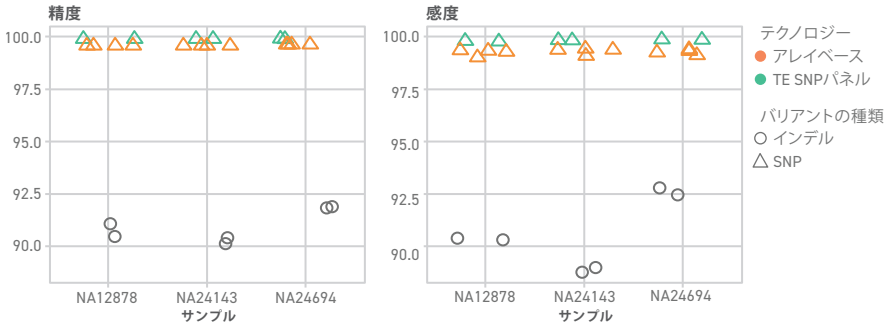


図 3：3 種類の異なる集団の GiAB サンプルにおける Twist SNP パネルと主要なジェノタイピングアレイの性能の比較。

平均カバレッジ	反復	10X (0.84 GB SEQ)	15X (1.4 GB SEQ)	18X (1.6 GB SEQ)	20X (1.9 GB SEQ)
精度	反復 -1	99.20	99.82	99.88	99.90
	反復 -2	99.20	99.80	99.86	99.89
感度	反復 -1	90.39	97.75	98.84	99.33
	反復 -2	90.50	97.72	98.70	Se

表 2：各平均ターゲットカバレッジにおける SNP ジェノタイピング（NA12878 を用いたデータ）。

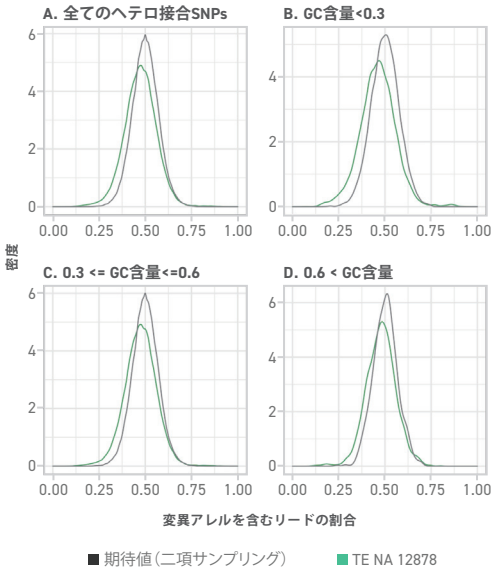


図 4：全てのヘテロ接合 SNP（A）または GC 含量で層別化した（B ～ D）、変異アレルを含むリードの割合。

結果を、NA12878 について個別に**表 1** に示し、バリアントの種類、テクノロジー、異なる集団からのサンプル別にまとめたものを**図 3** に示しました。SNP パネルの精度および感度は、アレイの場合と一致するかそれ以上であり、いずれも 99% 超でした。さらに、ジェノタイピングパネルにより、90% 超の精度と 88% 超の感度で挿入・欠失を特定することができました。

次に、NA12878 を用いてサブサンプリング解析を行い、平均 20x カバレッジまで指標が安定していることを確認しました。SNP 感度は 15x カバレッジ未満で著しく低下するものの、精度は >99% を維持します（**表 2**）。SNP 全体で所定の平均カバレッジを得るために必要なシーケンス量も表に示しました。1.9 Gb のシーケンスにより、約 25 万個の SNP のパネルについて高感度なジェノタイピングの利用が可能になります。SNP パネルの精度は 10x カバレッジでも 99% 超を維持すること（**表 2**）に注目すると、同量のシーケンスは 50 万個の SNP 以上のパネルを十分扱うことができ、ある程度の偽陰性を許容する、あるいは、インピュテーションや系統推定（ancestryinference）など SNP 全体にわたって脆弱な情報を統計的に統合するアプリケーションを実現できます。

アレルバイアスおよびコンテキストバイアスに対するジェノタイピング性能の堅牢性

製造コストを下げるため、ターゲットエンリッチメントパネルのプローブは通常、単一のゲノムシーケンスに対して設計されます。この結果、ゲノムリファレンスとは異なるアレルが含まれる遺伝子多型の位置でミスマッチが生じ、一方でそうでない場合は完全に一致します。設計を行う際にこのタイプの非対称性によってもたらされる可能性のあるバイアスを定量化するため、NA12878 について GiAB が報告している 70,693 箇所のヘテロ接合 SNP を GBS パネルにおいて観察したところ、非参照アレルを含むリードの割合は平均 47% であり、バイアスがない場合に予測される平均値 50% からの逸脱はわずか 3% でした（**図 4A**）。この小さなリファレンスバイアスは、バリアントを含むリードよりも、もともとリファレンスゲノムに完全に一致するリードとの相性が良く（Lunter および Goodson 2011、Degner ら 2015）リードアライメント・アルゴリズム単独でもたらされると知られているバイアスの範囲内であり、朗報といえます。GC 含量で層別化したところ（**図 4B ～ D**）、含量がとても低い場合（**<30%、図 4B**）プローブの平均値 46% に向かってわずかな増加が観察されました。これにより、キャプチャされた部位数が 6% 増加し（**図 4A、4B の緑色の曲線**）、両方のアレルについて完全に 50:50 の確率である場合の標本分散から予測される値を下回りました（**図 4 のすべての図におけるグレーの曲線。**実験方法を参照）。

ジェノタイピングおよび高カバレッジの全エクソーム解析

SNP パネルのキャプチャ性能を単独で試験することに加えて、当社の Core Exome およびコーディング配列と比較して平均 0.5 倍の SNP のカバレッジが得られるように量を設定したスパイクイン（エクソーム +0.5x SNP）を用いて、高カバレッジの全エクソームシーケンス（WES）と合わせて SNP ターゲットの複合キャプチャを行い、検証しました。**表 3** に、その複合キャプチャ実験に用いた各ターゲット、または組み合わせ（SNP のみ、エクソームのみ、または両方）について、150x の raw シーケンス後に得られた主要なターゲット指標を比較して示します。各パネルを個別に用いてキャプチャした場合に得られた指標と、GBS パネルにおいて SNP とプローブで異なるターゲットも比較のため示しています。

全てのターゲットにわたる解析（1 つの遺伝子座について**図 5** に示します）では、両パネルの主要なキャプチャ指標が複合キャプチャ後にも維持されることが明らかになりました。AT/GC のドロップアウト率および fold 80 は変化なく、平均カバレッジは SNP に対して選択した 0.5x カバレッジに応じて予想通りの反応を見せました。唯一の例外は、SNP ターゲットに対する fold 80 ベースペナルティがわずかに増加したことですが、値は依然として 1.4 を下回っており、ターゲットのキャプチャパネルに対して最高の均一性を示しています。設計上の理由で（片方のパネルの平均値はもう片方の約 0.5 倍）、複合パネルにおけるカバレッジ分布はもはや単峰型ではないことを前提とすると、複合キャプチャ実験においては Exome + SNP ターゲットに関する fold 80 ベースペナルティが決定されないことは注意するべき点です。

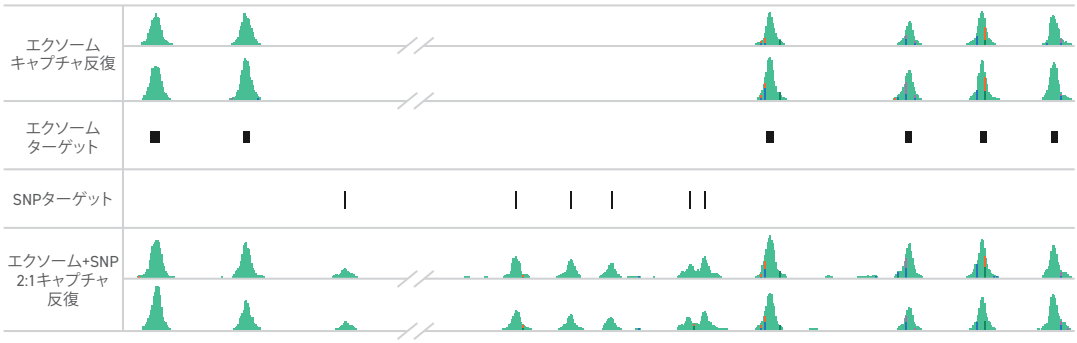


図 5：hg19 内の chr1 における 165,320,712 ～165,379,000 bp の小領域における 0.5X SNP、1X エクソームの複合キャプチャパネルに対するカバレッジ（165,330,000 ～165,362,000 bp の小さなセグメントを除くことで分割表示し、見やすくした）。それぞれのキャプチャピークは、左側にあるのは LIM ホメオボックス転写因子 1 アルファ（LMX1A 遺伝子）の最初の 2 つのエクソン、次に SNP ターゲットのカバレッジ、右側にあるのはレチノイド受容体 X ガンマ(RXRG)の最後の 3 つのエクソンに対応する。

キャプチャ実験	評価対象領域	平均カバレッジ	ターゲット塩基 20X (%)	ターゲット塩基 30X (%)	AT ドロップアウト	GC ドロップアウト	FOLD 80
エクソーム +0.5xSNP	エクソーム +SNP	71	98%	97%	2.26	0.68	—
	SNP のみ	37	94%	71%	0.53	0.01 未満	1.38
	エクソームのみ	72	98%	97%	2.24	0.53	1.37
エクソーム	エクソン	57	98%	95%	1.78	0.41	1.32
SNP	SNP 塩基に対して設計されたプローブ	44	98%	86%	4.48	0.05	1.35
	SNP	51	99%	95%	0.39	0.01 未満	1.29

表 3: SNP ターゲットとエクソームターゲットの両方からなる複合パネルのキャプチャ性能。

考察とまとめ

本研究においては、一般的なジェノタイピングアレイ内であらかじめ規定した一連の SNP を用いた、シーケンスによるジェノタイピング（GBS）のための Twist 独自の製造およびカスタムターゲットエンリッチメントパネル設計に関する性能を調べることに焦点を当てました。

全てのサンプルについて、並外れた均一性と低い Duplicate 率、SNP ターゲットの高カバレッジ、アレイベースのジェノタイピング以上のコール率という優れた性能を示す革新的な結果が得られました。SNP パネルによって、一段と広範囲にわたるインデルのような、より複雑なバリアントの発生を検出することもできました。

カスタム GBS パネルでは、Twist Core Exome パネルと比較してオフターゲットキャプチャの増加も認められましたが、プローブにより生成されるシーケンスカバレッジのピークと比較して SNP のターゲットプロファイルが狭いため、カバレッジ指標はエクソンのそれと一致するかそれ以上でした。ごく軽度のリファレンスバイアスが低 GC で顕著にみられましたが、カスタム GBS パネルは SNP 全体でわずか 20x の平均カバレッジで 99% を超える感度・精度のジェノタイピングを可能としました。さらに、SNP 全体で情報が統合されるジェノタイピングアプリケーション（インピュテーションや系統分岐など）のために、SNP 全体の平均カバレッジ 15x で 97% の感度、10x で 90% の感度が、>99% の精度とともに維持されることは特に注目に値します。

また、当社の高性能なパネルによって、NGS ベースの大規模なジェノタイピングを単独で実施することが可能になります。例としては、ゲノム系統分岐推定、インピュテーションによる全ゲノム関連研究、さまざまな深度のジェノタイピングアプリケーションが挙げられます。Twist はターゲットエンリッチメントを用いたワークフローとカスタム GBS パネルの優れた適合性により、様々なアレイ単独のアプリケーション、あるいはアレイベースのジェノタイピングをシーケンスと並行して行っている複合アプリケーションのために、ジェノタイピングアレイに代わる優れた選択肢を提供いたします。

参考文献

Degner J., et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. 2009. Bioinformatics, 25 (24) , 3207–3212 https://dx.doi.org/10.1093/bioinformatics/btp579

Heng L, Richard D. Fast and accurate short read alignment with Burrows-Wheeler transform. 2009. Bioinformatics, 25 (14) , 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Krusche, P, Trigg, L, Boutros, PC et al. Best practices for benchmarking germline small-variant calls in human genomes. 2019.Nat Biotechnol, 37, 555–560. https://doi.org/10.1038/s41587-019-0054-x

Lunter G., Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. 2011. Genome Res, 21 (6) , 936-9. https://doi.org/10.1101/gr.111120.110

Rayner NW, McCarthy MI. Development and Use of a Pipeline to Generate Strand and Position Information for Common Genotyping Chips. ASHG Conference Poster. https://www.well.ox.ac.uk/~wrayner/tools/

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. 2013. Curr Protoc Bioinformatics, 43 (1110) . https://doi.org/10.1002/0471250953.bi1110s43