Twist Human Pangenome Panel

スパイクイン方式のヒトパンゲノム

次世代シーケンシング（NGS）の普及に伴い、ヒトゲノミクスへの取組みは、全世界のヒトのバリアントの特徴をより深く明らかにすることに焦点が当てられるようになりました。この取組みと並行して、NGS アッセイで必要となる補助ツールも劇的に改善されてきました。Twist Bioscience では、NGS ベースのジェノタイピングパネル [1] がアレイに代わる選択肢として有効性を示すことや、集団バイアスのないゲノムワイドインピュテーションパネルの実用化 [2] など、NGS ソリューションの実現に注力してまいりました。

この分野において最も進歩したものの 1 つに、ヒトパンゲノム [3] が挙げられます。この新しいヒトリファレンスは、現在 49 の telomere-to-telomere（T2T）ヒトゲノムを組み込んでおり、今後数百にまで拡大する予定です。ヒトパンゲノムにより、複雑なバリアント、これまで明らかになっていなかった困難な領域、世界中の多様な祖先を新しいグラフ形式で提示することができます。さらに、パンゲノムデータを処理するための新しいアライメント方法や解析手法が開発されています。

パンゲノムベースのアッセイと開発を支援するために、当社は Twist Exome 2.0 パネル（hg38 ベース）の拡張版である、ヒトパンゲノムスパイクインパネルを設計し、本文書でその評価結果を紹介します。当社のシステムにおけるベイトのバリアント耐性に関する情報を活用し、カスタムプローブ設計戦略と組み合わせることで、コード配列と重なる新規のパンゲノムリファレンスのバリアント塩基の大部分をターゲットにすることができます（例として図 1A †を参照）。5 bp 以下のバリアントの場合、当社の既存のベイトにより 90% を超える有効性でカバーされます（図 1C を参照）。また、5 bp 超のバリアントに対しては、2.5 Mb の総スパイクインターゲットセットおよび 11.7 Mb のベイトフットプリントを含む、全パンゲノムバリアント塩基の 94% を効果的にカバーする新しいプローブを設計しました（図 1B を参照）。

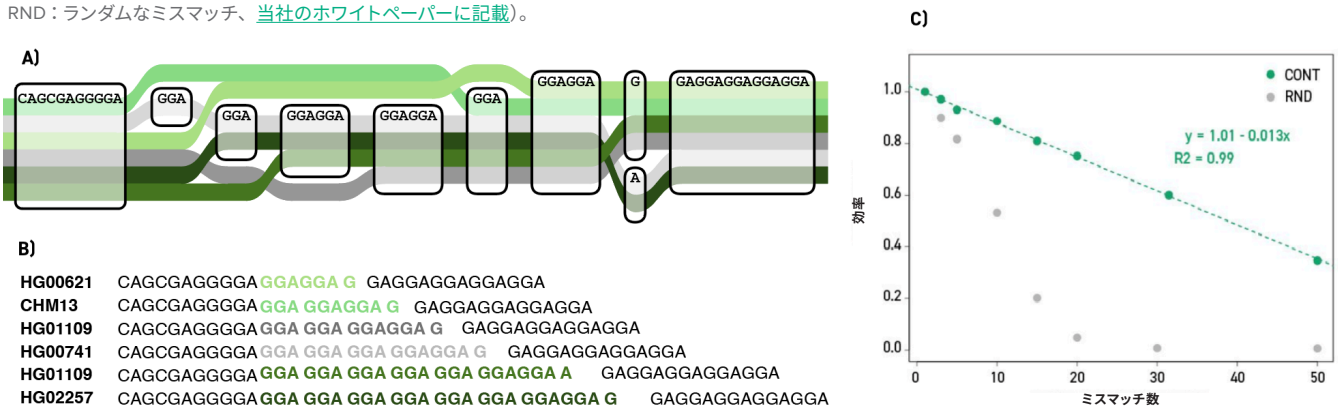
ワークフロー

スパイクインのパフォーマンスを評価し、当社のエクソームと比較するため、2 種類の細胞株を含むパンゲノムリファレンスセット（HG02257 & HG00261）で、エクソームパネル単独またはエクソーム + パンゲノムスパイクイン（Exome + PanG）を用いて反復キャプチャを行いました。キャプチャ後ライブラリは、75 bp ペアエンドリードの NextSeq 500 でシーケンシングし、各パネルの総ターゲット塩基に対して 150X までダウンサンプリングしました。次に、BWA[4] および CHM13 リファレンス [5]（v.2.0）を用いた標準的なアラインメントの結果と、Giraffe および CHM13 minigraph cactus パンゲノムリファレンス（vg v1.49.0 および hprc-chm13v.1.0）に基づくアラインメントの結果を比較しました [6]。

全体的なキャプチャ性能の結果

パンゲノムバリアントが hg38 リファレンスよりも T2T ゲノムにて難しい配列に富んでエンリッチされることが予想される一方、多くのバリアントが実際には低い複雑性を持っていることが示されました（図 2）。Exome 単体と比較し、Exome ＋ PanGenome では MAPQ フィルターは 2%、オフベイトは 3% と、増加率はわずかでした。また、パンゲノムベースの解析手法では、標準的な BWA と比較してオフベイトのわずかな改善が見られ、エクソーム単独対 Exome ＋ PanGenome の性能差を縮める結果となりました。その他の指標では、エクソーム単独と比べほぼ同等あるいは改善が見られたことから、検出困難なターゲットにもかかわらず、本パネルが有効であることが示されました。本パネルを用いることで、ターゲットに対するキャプチャ有効性への影響を最小限に抑えつつ、優れた総合パフォーマンスを達成することができます（例：エクソーム+ PanG とエクソーム単独を比較したところ、平均ターゲットカバレッジはそれぞれ 62 倍対 65 倍、30X 塩基%はそれぞれ 95% 対 96%）。

図 1. パンゲノムスパイクインデザイン。パンゲノムグラフ（パネル A †）で示されるバリアントと当社 hg38 エクソームに重複するターゲット領域は、hg38 に含まれないバリアント配列（パネル B）に対するベイト設計にそのまま使われています。当社のエクソームに対するベイトのカバレッジ、内容、重複は、キャプチャ有効性におけるミスマッチの影響に関する実験データに基づいて最適化されています（パネル C、CONT：連続したミスマッチ、RND：ランダムなミスマッチ、当社のホワイトペーパーに記載）。



† hg38 の chr20 pos 46022976 について、全 49 のハプロタイプのうち一部のみを示しており、発達および神経疾患と OMIM リンクを持つ SLC12S5 カリウム・クロライド共輸送体のバリアントとの重複領域を示しています。pangenome graph hprc-v1.1.-hg38、vg version 149.0 および SequenceTubemap を用いて図を作成しています。

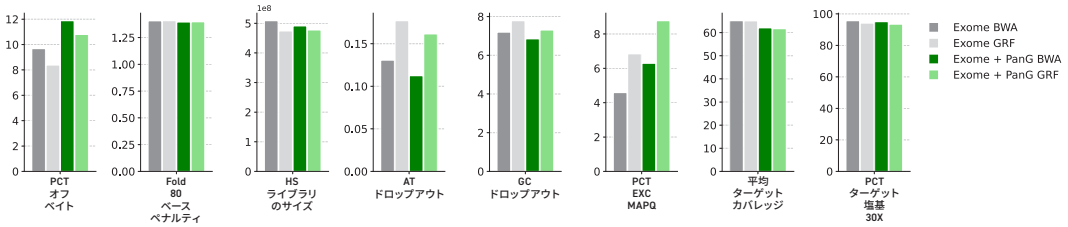


図 2. Picard キャプチャメトリクス。解析手法［標準的な BWA アラインメントと Giraffe を用いたパンゲノムベースアラインメント（GRF）の比較］およびパネル（エクソームとエクソーム＋パンゲノムスパイクインの比較）の効果を示しています。

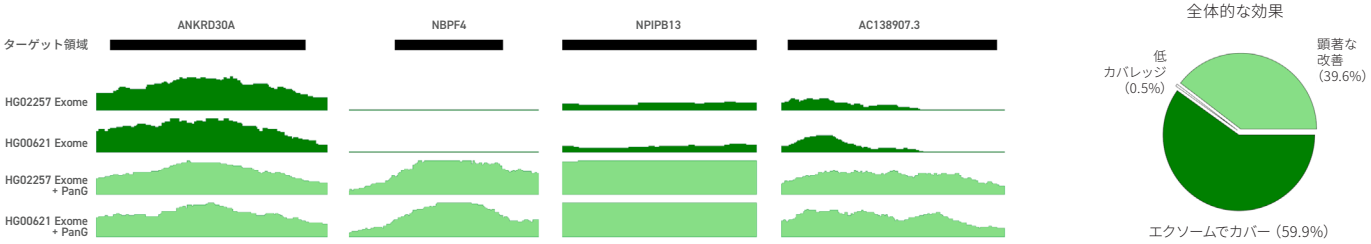


図 3. CHM13 ユニーク領域の 207 を超えるコード配列におけるスパイクインの効果。様々な細胞株でのキャプチャ結果に基づく、エクソーム単独（Exome）およびエクソーム＋パンゲノムスパイクイン（Exome + PanG）で観察されたカバレッジパターンおよび改善の例。一部の領域は両パネルで十分にカバーされていたものの（一番左）、他の領域ではパンゲノムスパイクインにより顕著な改善が示されました。右側の円グラフは、主なパターンの割合を示しています。1領域のみ平均カバレッジ 20X 以下（塩基間では 5-10X）を示しました。

パンゲノムパネル領域におけるカバレッジの改善

どのゲノムにおいても、コード配列上ではサイズの大きなバリアントの数は少ないため、様々な細胞株を用いなければエクソーム単独とパンゲノムのキャプチャ性能の差を統計的に比較することは困難です。実際、PanG で拡張した領域に対するジェノタイピング（Giraffe + deep variant[7,8] を使用）の結果、309 のバリアントしか検出されず、細胞株に 50 bp 以上のバリアントがあったにもかかわらず、エクソームパネルとエクソーム + PanG パネル間でこの 309 のバリアント検出がすべて一致しました。

しかし、主に T2T ゲノムの改善によって明らかになった一連の 207 のコード領域については、CHM13 と比較して hg38 では存在していないのに対して、パンゲノムではすべてのサンプルで解析されています。したがって、この CHM13 ユニーク領域 [5] は、hg38 には存在しない、パンゲノムにおける相当数のバリアント解析の優れたモデルとなります。

CHM13 ユニークのコーディング領域の大部分が、Twist Exome 2.0 パネルで十分にカバーされているか（約 60%：エクソーム単独でも多くのパンゲノムバリアントをキャプチャしていることを裏付けています、序文参照）、スパイクインでより顕著に改善していました（約 40%）（図 3）。両パネルともに平均カバレッジが 20X 未満の領域は 1 つのみでした。パンゲノムスパイクインを用いることで、追加の 1.4% のゲノム領域と 6.5% の全コーディング領域（CHM13 ユニーク領域を除く）にて統計的に有意な改善がみられており、エクソーム単独の場合の平均カバレッジ四分位範囲と比べそれぞれ 1.5 倍、および 3 倍を超える差が観察されました。なお、パンゲノムスパイクインおよびパンゲノムのアラインメント手法で扱う、より複雑なマッピング環境では、統計学的に有意な性能低下は認められませんでした。

結論

パンゲノムで認識されるエクソームでターゲットとされるバリアントの検出困難な性質にもかかわらず、パンゲノムバリアント領域でターゲットエンリッチメントパフォーマンスが著しく改善し、パネルのその他のパフォーマンスが著しく低下することもありませんでした。

パンゲノムスパイクインにより、大規模で複雑な集団情報に基づくバリエーションのキャプチャが明確に改善しました。

今回開発および検証されたフルセットのパンゲノムバリアントを含んだ方法とアプローチは、他の方法にも容易に適用でき、難しいターゲット領域に特化してパンゲノムバリアントを注意深く特徴付ける場合や、他の生物種にも適応できます（パンゲノムは複雑なゲノムや農業に関する重要な品種を解釈する上で強力なツールとなります）。また、ターゲットロングリードシークエンスなどを含む他のシークエンス技術やアプリケーションに応用できます。

パンゲノムターゲットエンリッチメントパネルに関する詳しい情報は、当社までお問い合わせください。

twistbioscience.com/ngs
sales@twistbioscience.com

参考文献

1. Capture-based SNP Genotyping with Twist Target Enrichment Panels. Twist Bioscience https://www.twistbioscience.com/resources/application-note/capture-based-snp-genotyping-twist-target-enrichment-panels (2020).

2. Abecasis, G. No More Arrays: Genotyping by Sequencing Enables Economical and Improved Association Studies. Twist Bioscience https://www.twistbioscience.com/resources/webinar/no-more-arrays-genotyping-sequencing-enables-economical-and-improved-association (2021)

3. Liao, WW., Asri, M., Ebler, J. et al. A draft human pangenome reference. Nature 617, 312–324 (2023). https://doi.org/10.1038/s41586-023-05896-x

4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009). https://doi.org/10.1093/bioinformatics/btp324

5. Nurk, S. et al. The complete sequence of a human genome. Science 376, 44–53 (2022). https://doi.org/10.1126/science.abj6987

6. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science 374, abg8871 (2021).https://doi.org/10.1126/science.abg8871

7. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. 36, 983–987 (2018). https://doi.org/10.1038/nbt.4235

8. Rakocevic, G. et al. Fast and accurate genomic analyses using genome graphs. Nat. Genet. 51, 354–362 (2019). https://doi.org/10.1038/s41588-018-0316-4