

法医学ゲノミクスにおける ターゲットシーケンスの価値

次世代シーケンサーテクノロジーでは、1日でヒト全ゲノムのシーケンスを行うことが可能になりました。全ゲノムシーケンス（WGS）は個体間の遺伝暗号の違いのすべて（コーディング、調節およびイントロン領域に生じるものを含みます）へのアクセスを提供し、疾患や生体系の研究に有用です。全ゲノムシーケンスは最高のゲノムカバレッジを提供する一方で、シーケンスと解析に多大な労力を要します。法科学者はゲノムの広い範囲のデータを得るより、法医学 PCR 産物のターゲットシーケンスの実施を選択します。法医学分野で利用される高密度の遺伝子座セットのシーケンスを行うことによって、ケースワークとデータベースにかかる労力を法医学的な疑問に最良の答えをもたらすゲノム領域に向けられます。このことはプライバシーの問題を軽減するだけでなく、全ゲノムシーケンスよりも少ないデータですむことから、既存の法医学 DNA ワークフローにおいて一般的なボトルネックとなっている解析の単純化につながります。次世代シーケンサーはアリルをサイズごとに分離しないため、より幅広い答えをもたらす能力が著しく高まっています。

キャピラリー電気泳動の限界を超える

次世代シーケンサーは（1）1種類のワークフローを利用して1回の反応で複数の遺伝子多型を解析すること、（2）現在用いられているマーカーに関するより高分解能のジェノタイピングを行えること、（3）分解された DNA サンプルから有用なゲノム情報を最大限に引き出すこと、（4）より高分解能の mtDNA および混合サンプルの解析を可能にすることによってキャピラリー電気泳動の限界を超えて、法生物学を改善する可能性があります。

複数の遺伝子多型の同時解析

キャピラリー電気泳動の場合、異なる遺伝子多型を同時に解析できないため、法医学研究室では PCR を用いる複数のシステムを検証および維持する必要が生じます。異なるクラスの遺伝子多型を解析するには複数の手順を組み合わせなければならず、それぞれの手法に独自の QA/QC およびトレーニングプログラムが必要です。たとえば、多アリル常染色体 STR を区別するためにキャピラリー電気泳動による断片サイジングを行う場合、一般的には Y または X 染色体マーカーに関する同様の解析が別々に行われ、これらの手順のすべては一塩基多型（SNP）同定のための 1 塩基伸長アッセイのワークフローとは異なります。サンガー法による mtDNA のシーケンスを行うにはさらに別のシステムを要します。この結果、解析が困難な DNA サンプルを複数回調べ、全データの抽出を試みる必要が生じることもあります。

アプローチが互いに異なるため、解析担当者はどの方法を行うかについての意思決定が必要です（図 2）。少数の STR にフォーカスすべきでしょうか？ それとも mtDNA でしょうか？ SNP でしょうか？ 少量または低品質のサンプルを用いて異なるテクノロジーにより複数のアッセイを行わなければならないこともあります。特に、非常に損傷のある DNA サンプルや少量の DNA サンプルを取り扱う場合には、1つのアッセイを選択することで別のアッセイが実施できなくなる可能性もあります。完全なデータセットが得られなければ、部分的な結果や確定的でない結果しかもたらさないため、このようなサンプルは未処理のままです。

次世代シーケンサーはバーコードを付加することによってサンプルのマルチプレックス処理を可能にし、従来のキャピラリー電気泳動法では不可能なレベルでの優れたスループットでデータベース構築をサポートします。既知の DNA バーコード（インデックス配列）を使うことによって、法医学サンプルを制御下でプール（混合）し、数十から数百サンプルのシーケンスを同時に行うことができ、各サンプルはデータ解析中に自動識別されます。またインデックス化は内部 QC 機能も提供し、インデックスタグがサンプルに付加されると、これらはサンプル処理および保管のすべてのフェーズで利用可能な追跡要素となります。

次世代シーケンサーはゲノム全域から遺伝子座のデータを提供することができます。なぜそうする必要があるのでしょか？ 本当のところは、それが既にそういうものだからなのかもしれません。DNA データベースに登録するために処理されるサンプルの数が増加し、管轄区域同士が各データベースの情報を共有し始めたために、複数の国際的データベースは 1 回のランで解析する遺伝子座を増やす必要性が出てきたため、遺伝子座データベースの定義を拡張しました。拡張された遺伝子座セットは、容疑者が同定されるまたはデータベース検索でヒットする場合に法執行捜査の効率を改善することになるでしょう。容疑者がおらず、データベース検索でヒットもしない場合には、証拠サンプルから得られる追加の情報が、収監されておらず再犯の危険性があるかもしれない犯罪者の表現型に関して貴重な手がかりをもたらす可能性があります。

高分解能ジェノタイピング

次世代シーケンサーは、同じ長さで遺伝子配列が異なるアリルをキャピラリー電気泳動による検出に基づいて識別する能力も兼ね備えています。長さのみに基づいて低分解能ジェノタイピングを行うキャピラリー電気泳動は STR 内に存在する SNP を検出することができません（図 3）。ケースワーク比較において、特に部分的に分解されたサンプルを用いて最大限の関連データを引き出さなければならぬ場合には、配列変異が重要となることがあります。次世代シーケンサーを利用すれば、犯罪科学者は最も完全な結論に到達することができます。

図 3：次世代シーケンサーを用いた STR 内 SNP 検出



次世代シーケンサーは配列変異のある同じ大きさのアリル（D31358 アリル 16 と SNP が存在する D31358 アリル 16）を検出できます。キャピラリー電気泳動は STR の長さのみに基づいてジェノタイピングを行うため、これらの STR 内アリルを同定することはできません。

分解されたサンプルから情報を最大限に引き出す

キャピラリー電気泳動の限界は、鍵となる遺伝子座の存在の検出能にも影響を及ぼします。キャピラリー電気泳動は少数の STR アリルのサイズ比較に利用できますが、5 または 6 色の色素を用いるキャピラリー電気泳動装置で遺伝子座を明確に識別するためには、サイズ範囲が重複しないように遺伝子座を設計しなければなりません。これは、キャピラリー電気泳動解析のために一部のアリルを 450 塩基対以上に伸長しなければならないことを意味します。たとえば犯行現場サンプルや遺体から採取した DNA のように DNA が分解されている場合には、長い断片が検出されず、情報量が少ない不完全な結果しか得られない可能性もあります。

次世代シーケンサーは法医学 PCR アリル断片検出における「小さい」という用語を再定義します。イルミナの SBS テクノロジーを使えば、同時に解析可能な遺伝子座の数やアリルの長さの限界はありません。このテクノロジーは、キャピラリー電気泳動を用いた長さに基づくジェノタイピングで直面する「限られた物的財産」の問題を解決します。この結果、次世代シーケンサーを用いた法医学ゲノミクスアプローチは、証拠サンプルや既知の参照サンプルから最大限の情報を引き出すことができます。イルミナの SBS は、小さなマーカ（75 塩基未満）からなる高密度のセットのジェノタイピングによって分解された DNA サンプルの解析を改善するので、可能な限り多くの遺伝情報を引き出すことができます。

高分解能 mtDNA および混合解析

深いシーケンスカバレッジは、マイナーな DNA 要素や微量の DNA 混合物の低濃度検出と定量を容易にします。キャピラリー電気泳動法ではこれらは無視されるか、一部しか検出されません。カバレッジとは、ある塩基が、1 回のシーケンスランで何回読まれたかを意味します。たとえば、30x カバレッジでシーケンスされた法医学アンプリコンのセットとは、各塩基が平均して 30 のシーケンスリードによりカバーされることを意味します。

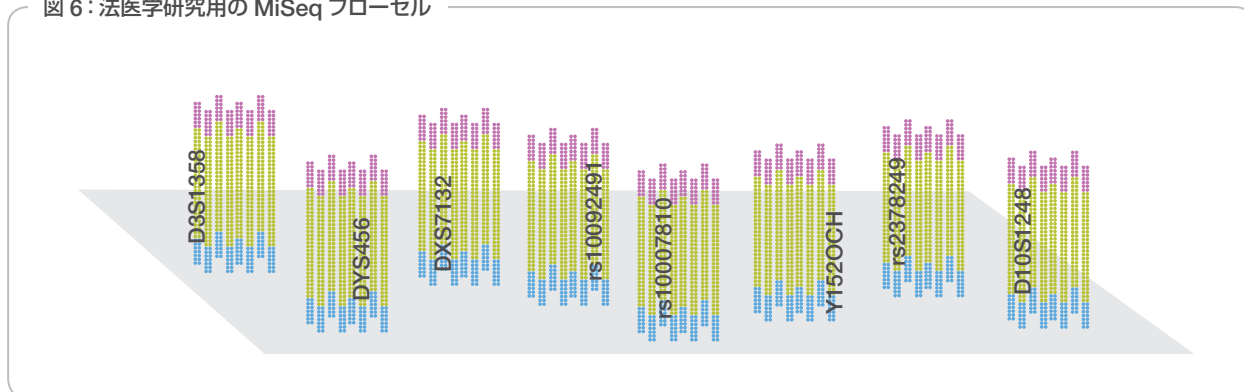
カバレッジは mtDNA を解析する際に特に重要であると考えられます。ヒトの細胞には、数千もの mtDNA コピーが含まれていますが、各コピーは必ずしも一致するものではありません。ヘテロプラスミーと呼ばれるこのミトコンドリア変異は、法医学サンプルの比較を難しくすることがあります。キャピラリー電気泳動を使ったサンガーシーケンスでは、マイナーな変異の頻度がメジャータイプの 10～20% の場合、こうしたヘテロプラスミーを精度よく検出できます。しかし、イルミナの SBS ケミストリーがもたらすより深いリードは、変異頻度が約 1% の場合でも高精度にヘテロプラスミーを検出することを可能にします。このため、より明確なヘテロプラスミー解析ができるだけでなく、DNA 混合物中における個々の要素のより高解像度な解析を実施することができます。

次世代シーケンサーが提供する精度の高いデジタル式のリードカウントは、明確なデジタルシグナルを通じたヌクレオチド配列の直接観察を可能にし、「ノイズ」という用語を再定義します。次世代シーケンサーを使って得られるリードはヒト DNA であり、ピークの色やサイズ、高さ（相対蛍光単位 [RFU] として）そして形を主に測定するアナログのキャピラリー電気泳動システムとは対照的です。次世代シーケンサーが提供するデジタル式のリードカウントは、サンプル中に存在する各変異をはっきりと「見る」能力を高め、DNA 混合物に関するより多くの情報をもたらします。たとえば、次世代シーケンサーのデジタルシグナルは mtDNA 解析を 1 塩基ごとに明らかにします。これとは対照的に、サンガーシーケンス法ではあいまいなベースコール（例：不完全なスペクトル分離、塩基の誤取り込み、一様でないベースラインなど）を生じ、無駄なデータが得られます。

TruSeq[®] シーケンスケミストリーは、SBS のプロセスをフローセル固相表面上に形成された数百万のクラスターにおける同時反応の大量並列工程を行います (図 6)。

ランが開始されると、MiSeq はクラスター形成と、その後にシーケンスを実施します。ペアエンドケミストリーは、フローセル上に形成された各クラスターのフォワードおよびリバーステンプレート鎖の読み取りに利用することができ、配列をより確実なものにします。2014 年 1 月現在、法科学者および法医学研究者は 500 サイクル用の MiSeq キットを用いて最大 525 サイクルを実施でき、1 回の反応ですべての法医学的遺伝子座をカバーするリード長をもたらします。

図 6: 法医学研究用の MiSeq フローセル



法医学ゲノミクスにおいて、次世代シーケンサーのデータの品質や精度は、特に混合 DNA サンプル、mtDNA ヘテロプラスミーや SNP データに関する結果を報告する場合に重要です。次世代シーケンサーのデータ精度は、特定の位置において塩基が正確にコールされる確率を測定し、指数として表されます。この値は Phred やクオリティ (Q) スコアとして知られており、範囲は 4 ~ 60 で、値が高いほどベースコールの品質が高いことを示します。たとえば、Q10 はベースコールが正しくない確率が 10 分の 1 であること、あるいは 90% の確率でそのコールが正しいことを示し、Q30 はベースコールが正しくない確率が 1000 分の 1 であること、あるいは 99.9% の確率でそのコールが正しいことを示しています。TruSeq ケミストリーで強化されたイルミナの SBS は、エラーフリーのリード率が世界で最も高く、Q30 を超えるベースコールを最も多く提供します。

データ解析は MiSeq 装置上で実施することができます (付属 PC は必要ありません)。ベースコールが完了した後で行われる二次解析には、適切なリファレンスゲノムへのアライメント (例: ヒトゲノム DNA または mtDNA、revised Cambridge Reference Sequence [rCRS]) を使用) およびその後のアليلコールが含まれます。

イルミナでは現在、ライフサイエンスアプリケーション向けに構築された使いやすいソフトウェアを発展させた、特殊な法医学アプリケーションのニーズを満たすパイオインフォマティクスソフトウェアを開発中です。シンプルなワークフローは、統合的なデータ解析の迅速かつ容易な実施を可能にします。

ライフサイエンスアプリケーション向けの全ゲノムシーケンスは数テラバイトのデータを産出しますが、特定のゲノム領域 (300 アンプリコン以下) にフォーカスする法医学的次世代シーケンス解析アプリケーションの産出データ量はこれよりもはるかに少なくなります。その結果、法医学的次世代シーケンス解析アプリケーションに必要なデータストレージ容量は、全ゲノムシーケンスよりも、キャピラリー電気泳動を用いるシーケンス法に近いものとなっています。イルミナは安全なデータ共有と保管のためのクラウドソリューションを既に提供しており、法医学ゲノミクス用の追加ツールや推奨事項についても開発中です。

次世代シーケンサーデータと既存のデータベースの互換性

世界中の多くの管轄区域内において、法執行および政府機関は DNA データベースの使用を通じて公共の安全性を高め、無実の人の容疑を晴らし、行方不明者の同定に協力しています。次世代シーケンサーのアリルコールはこれらのデータベースと互換性があり、新旧データのスムーズな連携を可能にします。次世代シーケンサーのアリルコールはさらに多くのものを提供します。ある STR 全体を直接シーケンスすることによって、次世代シーケンサーは 1 塩基ごとの配列と、反復単位全体の数だけでなく、不完全な反復および変異アリルの数も示します。たとえば、テトラヌクレオチド STR の場合には $\pm 0.5\text{bp}$ のサイズ区間に納まるのではなく、1、2 または 3 ヌクレオチドの付加または欠失の差を有するアリルが直接、明らかにされます。既存の法医学的アリルの命名法はそのままで変わっていません。基本的なアリル指定データに加えて、次世代シーケンサーは反復配列内データの力を利用して、同じ長さで内部配列の変異を有するアリルを持つサンプルの分解能を高めます。

結論

法医学サンプルの解析に関して、次世代シーケンサーがキャピラリー電気泳動によるアプローチを超える多くの利点をもたらすことは明らかです。ヒトゲノムプロジェクト以前の技術を利用する現在の法医学ワークフローは、ゲノミクスの力をわざと切り捨てており、より完全な遺伝子プロファイルを生成するためには解析を複数回行う必要があります。これとは対照的に、次世代シーケンサーテクノロジーは、法科学者が微量のサンプルや損傷があるサンプルから入手した場合であっても多くのヒト遺伝情報を活用することを可能にし、結果の品質と捜査の手がかりとしての有用性を高めています。

最良の次世代シーケンサーシステムは、ライブラリー調製プロトコールにシーケンステクノロジーを組み合わせ、法医学アプリケーションの厳しい要求に応じるために必要なサンプルサイズ、使いやすさ、カバレッジと精度の要件を満たしたものです（表 3）。

表 3：法医学アプリケーション向け次世代シーケンサーシステムの特長

項目	次世代シーケンサーシステムの特長
遺伝子座マルチプレックス能力	1 回のランで常染色体 STR 解析、Y STR 解析、X STR 解析および複数のタイプの SNP 解析を実施；数十から数百遺伝子座
サンプルマルチプレックス能力	1 ~ 384 サンプル
深いカバレッジ	ランあたり 1400 万リード
高品質データ	Q30 以上
正確な低濃度混合物検出	1% 超のマイナーな要素を検出
短いアンプリコンの検出	75 塩基未満
シンプルなライブラリー調製	90 分のプロトコール；ハンズオンタイムは 15 分未満
小さいサンプルサイズ	1ng 未満の DNA インプットサンプル

今後の展望

従来の法医学 DNA タイピングは、特に現在のキャピラリー電気泳動を用いるシステムの技術的限界と能力のギャップに対処する場合には、困難で長時間を要することが多くなりえます。この状況が法医学研究室の高まるニーズと組み合わせられると、研究室のインフラストラクチャーと解析担当者の双方への圧力と期待が大きくなります。包括的な法医学ゲノミクスは本当の答えを得るまでにかかる時間を短縮することができ、より多くのサンプルからより完全な法医学プロファイルを生成します。法科学に利用できる遺伝子座の数がこれまでよりも増加したため、ケースワークサンプルに関して多くの疑問に答えて情報量の多いデータを得ることが一層進み、捜査の手がかりをサポートします。しかし、まだ多くの作業を行う必要があります。公共の安全性を高め、犯罪者、行方不明者および災害時の大規模調査の感情的ストレスを軽減するために、新たなアプローチの開発と実用化が続けられています。

イルミナは、これらを達成しようと努力している法科学者の役に立つために全力を尽くしています。実証され広く利用されているイルミナナの SBS テクノロジーは、シーケンスを通じて変異を迅速かつ正確に同定するための信頼性の高いソリューションを提供し、最良の技術を学術研究の世界から DNA を解析する法医学研究室に移行させています。

参考文献

1. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et. al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30: 434-439.

AGAAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAAAGTATCAATTGAGACTAAATATTAACGTACCATTAAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATT
TCAACGTACCGTAACGAAAGTATCAATTGAGACTAAATATTAACGTACCATTAAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAAAGTATCAATTGAGACTAAATATTAACGTACCATTAAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATT
CGAAGGAAAGAAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAAAGTATCAATTGAGACTAAATATTAACGTACCATTAAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATT
CAACGTACCGTAACGAAAGTATCAATTGAGACTAAATATTAACGTACCATTAAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAAAGTATCAATTGAGACTAAATATTAACGTACCATTAAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATT
AGAAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAAAGTATCAATTGAGACTAAATATTAACGTACCATTAAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATT
GATTACTTGATCCACTGATTCAACGTACCGTAACGAAAGTATCAATTGAGACTAAATATTAACGTACCATTAAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAAAGTATCAATTGAGACTAAATATTAACGTACCATTAAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATT
CGTATCAATTGAGACTAAATATTAACGTACCATTAAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAAAGTATCAATTGAGACTAAATATTAACGTACCATTAAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATT

