論 文 の 内 容 の 要 旨

論文題目　**大規模計算システムを用いたがん体細胞変異検出アルゴリズムの研究**

氏　　名　　上田　宏生

がんとは制御されない細胞の増殖で、発生部位から他の部位への浸潤と拡散を示す病気の一群をいうが、一つの病気ではなく 100 種類以上のタイプのがんが分類されている。がん細胞のゲノムは体細胞点変異、欠損や染色体転座といった多くの変化を含んでおり、それら変異の蓄積によって発病することから、がんは「Disease of genome(ゲノムの疾患)」とも呼ばれている。よって、がん研究においてがんゲノムの変異を解析することは、がん研究の基盤として極めて重要である。また、近年、次世代シーケンサの技術進展は目覚ましく、がん細胞のゲノムに生じた体細胞変異を網羅的に検出することが可能になってきた。そこで、国際がんゲノムコンソーシアム（ICGC ：International Cancer Genome Consortium）やがんゲノムアトラスプロジェクト（TCGA: The Cancer Genome *Atlas*）において大規模ながんゲノム解析が進行している。これらのプロジェクトでは遺伝子上の体細胞変異を検出する方法として、エクソン特異的にＤＮＡを抽出し、次世代シーケンサを用いて全エキソンを網羅的に解析する全エキソームシーケンスと呼ばれる方法が中心的に用いられている。

全エキソームシーケンス方を用いることで、ゲノムワイドにがんに特異的な遺伝子上の体細胞変異の検出を行うことができる。しかし、エクソームシーケンスデータを含む次世代シーケンサデータ解析の問題点として、産出されるデータ量が膨大で解析に非常に時間がかかることが挙げられる。また、がんゲノムの解析においては、腫瘍細胞は純度（腫瘍率）の高い状態で検体が取られることが難しく、検体によって腫瘍率が均一でないことや、腫瘍におけるゲノムコピー数変異および、がん細胞のヘテロジェネイティ(不均一性)の存在がデータの中のノイズを増やす原因となっており、解析を難しいものとしている。

そこで、本研究では大規模な計算環境で効率的な並列処理を行うことにより高速に解析を行う方法を開発するとともに、がん細胞における体細胞変異を検出する為に、ゲノムコピー数変異と腫瘍率を算出し、それらの値でアリル頻度とリードデプスを補正することによりノイズ成分と真の候補を分離し精度の高い体細胞変異の検出を行う方法を開発した。まず、高速化のために、大規模計算環境での分散処理を自動化させる

必要があった。そこで、分散システム内で実装されている資源管理 API と協調して、計算機資源を有効に利用して処理時間を短縮して複雑な次世代シーケンサリード解析のすべてのプロセスの並列分散処理を行うことのできるフレームワーク AGORA framework を作成した。このフレームワークはオープングリッドフォーラムによる標準規格の Distributed Resource Management Application API (DRMAA) を用いて実装され、幅広い分散環境で動作することが特徴である。また、次世代シーケンサリードの標準フォーマットである fastaq および bam 形式でファイルの入出力を行うことから既存のプログラムを用いて容易に並列処理が行えるようになった。このフレームワークでは、xml (Extensible Markup Language) ファイルに処理フローを記述することにより、次世代シーケンサリード解析における、リファレンスゲノムへマッピング, ゲノム位置でリードの並べ替え、リアラインメント、体細胞変異候補の抽出といった汎用的な処理に対して自動化されたフローを構築できる。これにより次世代シーケンサの大規模データの処理が 30 倍程度高速化され、かつ容易になり、今日エクソームシーケンスをデータ量として標準的な 1 サンプルにつき 10～20 ギガベース(GB)のデータを 3 時間程度で解析できるようになった。

次に、この並列分散処理フレームワークを用いて、肝がんを中心とするエキソームシーケンスのデータ解析を行った。また、がん体細胞変異検出精度を向上させるためのアルゴリズムの開発を行った。このアルゴリズムが検出するのは、がんにおける体細胞コピー数変異、点体細胞変異、挿入、欠失および腫瘍率である。

本研究の方法では、まず、コピー数の絶対定量と腫瘍率の検出を行うアルゴリズムを開発した。このアルゴリズムでは、がんサンプルおよび非がん部のシーケンスリード数をエキソームのキャプチャーターゲットごとに集計し,がんと非がん部のリード数の比を集計する。次に連続ウェーブレット変換とマルチステイト隠れマルコフモデル(HMM)を用いて各コピー数のセグメンテーションを行い、各倍数性(ploidy)のピークとゲノム領域を関連づけた後、各ploidyピーク間の距離と該当するゲノム領域のヘテロSNPのアリル頻度から、それぞれのピークのアレル頻度とピーク間距離の特徴量を抽出し、理論的に予測されるploidyピーク間距離と、アレル頻度のインバランスと、腫瘍率からなるマトリックスとのベクトルマッチングを行うことにより、正確にコピー数変異と腫瘍率を検出する。次に、体細胞点変異検出アルゴリズムを開発した。エキソームシーケンスにおいては非がん部のリードが十分に取れないことやリードのマッピングエラーやシーケンスエラーが体細胞点変異検出における偽陽性候補の原因となる。これらを取り除く為に、従来使用されてきたフィッシャー検定とベイジアン確率によるフィルタリングを組み合わせるとともに、候補部位のリードを精査して候補領域の配列のエントロピーや、複数のホモロジーリードの混在、および周辺のミスマッチ塩基を精査し、また、公共の一塩基多形 (SNP) データベースのdbSNPの登録状況と他のノーマルサンプルでのパネルと比較することにより偽陽性候補を取り除くアルゴリズムとした。さらに、低リード深

度かつ低アリル頻度領域に、マッピングエラーやシーケンスエラー由来のノイズを十分に取り除く為に、ヘテロSNPの分布からノイズピークの存在範囲を予測し、腫瘍率補正後のアリル頻度およびリード深度からなる体細胞変異候補に対して、実測値のノイズピークを期待値最大化法（EMアルゴリズム）を用いてフィッティングさせることにより、ノイズピークの大部分を分離する手法を開発した。次世代シーケンサリードの解析においてはシグナルノイズ比は,それぞれのサンプルや実験、またリード数の総量に左右されるので、本手法は既存のソフトウェアの静的な閾値によるノイズ除去に比べて、より広範囲な条件で性能を発揮することが期待される。本研究で開発したアルゴリズムを用いて肝癌200例のエクソームデータで試験を行い、既存のアルゴリズムと比較することで体細胞におけるCNV,SNVおよび腫瘍率の検出を行ったところ、他の手法との比較において高い性能を有することが認められた。本手法は特に腫瘍率が低くノイズの多いサンプルや、コピー数変化が複雑ながんサンプルにおいて、エクソームシーケンスの情報からがんの原因遺伝子の特定をする場合に有用であると思われる。