大規模データ解析と人工知能技術による がんの起源と多様性の解明

伊東　聰, 角田将典, 田中洋子, 宮野　悟　（東京医科歯科大学　M&Dデータ科学センター）
{sito, mkakuta, hiroko-t, miyano}.dsc@tmd.ac.jp

はじめに

がんは今や死因の第一位であり人類にとって最大の脅威である。過去10年間におけるがんのゲノム解析によって、主要ながん種のほとんどについて、その病態に関わるドライバー変異の全体像が明らかにされた。それらの成果により、変異の原因や腫瘍内の多様性、病期の進展・再発が遺伝子変異の観点から理解できるようになった。また、最近、一見正常に見える組織においてすでにがんで特徴的に認められる遺伝子の変異を有するクローンの拡大が頻繁に生じていることが確認され、がんの前駆病変として注目を集めている。

一方、発がん初期のクローン選択の過程や、その後、多数の変異の獲得とクローン選択によってがん細胞集団に高度な多様性が生じ、浸潤・転移・再発が惹起される過程の分子メカニズムについては、なお多くが不明であり、がんゲノミクス研究の世界の中心となっている。本研究開発は、これらを理解するために、正常組織においてどのように遺伝子変異クローンが生じるのか、遺伝子変異ないしその組み合わせがどのように細胞の表現型を決定するのか、さらには、その多様性・複雑性のために研究が進んでいないゲノムの構造異常が発がんにどう関わるのか、について解明することを目的とする。

遺伝子解析とスーパーコンピュータ

がんの研究には遺伝子解析（全ゲノム解析）が必須である。全ゲノム解析は3つの処理で構成されている。まずはじめにサンプルからDNAを読み取るシーケンス、次に読み取ったDNA情報を数理的・統計的に処理する情報処理、最後に結果から医学的知見を読み取る結果解釈の3つである。2007年頃に登場した次世代シーケンサーは2021年現在に至るまで驚異的な性能向上を達成しており（図1）、一人の全ゲノムシーケンス（Whole Genome Sequencing: WGS）におよそ2日、600ドル程度となっている。

一方で、現在問題になっているのは2つ目の情報処理と最後の結果解釈である。次世代シーケンサーによって読み取られたDNA情報は1サンプルあたり数TB規模になっている。また、がんは時間的に多様な進化を遂げることがわかっており（図2）、そのメカニズムを解明するためには3000サンプル以上の解析が必要と言われている。また、遺伝子解析を医療の現場で用いるクリニカル・シーケンスにも注目が集まっている。日本の新規がん罹患者全員に全ゲノム解析を行うためには1日2000サンプル以上の解析が必要となる。この膨大なデータを処理するためにはスーパーコンピュータとその性能を活用できる解析ソフトウェアが必須である。当課題ではスーパーコンピュータ「富岳」上で高速に動作するヒト全ゲノム解析ソフトウェアGENOMONを開発している。

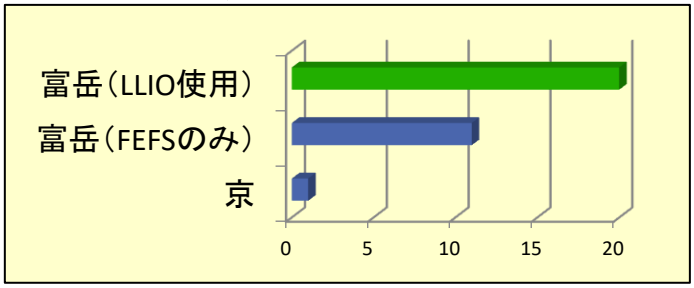
最後の結果解釈について、これまでは部分的にコンピュータ処理したデータから医師および生物学者が知識と経験に基づいて人海戦術で処理していたが、世界で初めて結果の説明を可能とするAIソフトウェアDeepTensorを富士通研究所と共同開発した（詳細は別紙ポスター）。

ヒト全ゲノム解析ソフトウェア「GENOMON」

スーパーコンピュータ「富岳」のCPUが採用しているArmv8.2-A SVEは、DNA解析ソフトウェアで最重要な整数演算を高速に実行可能である。また、他のスーパーコンピュータ用アプリケーションとの大きな違いとしてファイルI/Oが頻繁に行われる点が上げられる。そのため、アプリケーションの実行性能はCPU性能と並んでファイルシステムの性能に大きく左右される点が特徴である。

富岳は階層化ストレージを採用（図3）しており、京でも採用されたFEFSファイルシステム（第2階層）に加え、計算ノードからの高速アクセスが可能なLightweight Layered IO-Accelerator（LLIO）を第1階層に持つ。LLIOのディスク領域はFEFSに比べ小さいため、限られた領域にどのようなファイルを配置するかが性能を左右する。ファイル配置を含む全ての最適化を実施した結果、京の約20倍の高速化を達成（表1）。富岳全系を使用した場合、約6000検体/日の解析性能となった。

表1　GENOMONの性能向上



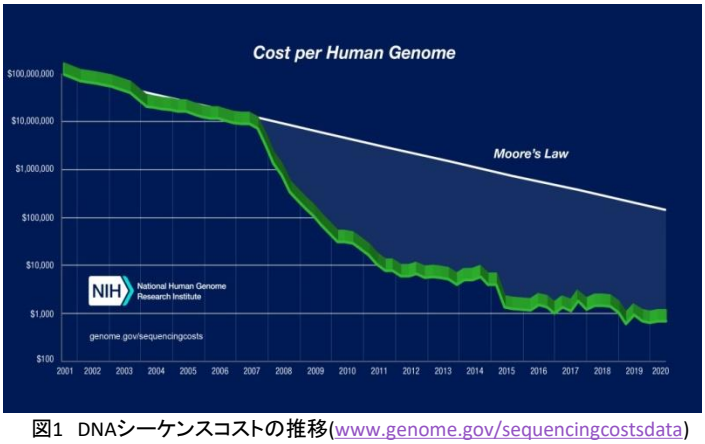


図1　DNAシーケンスコストの推移(www.genome.gov/sequencingcostsdata)

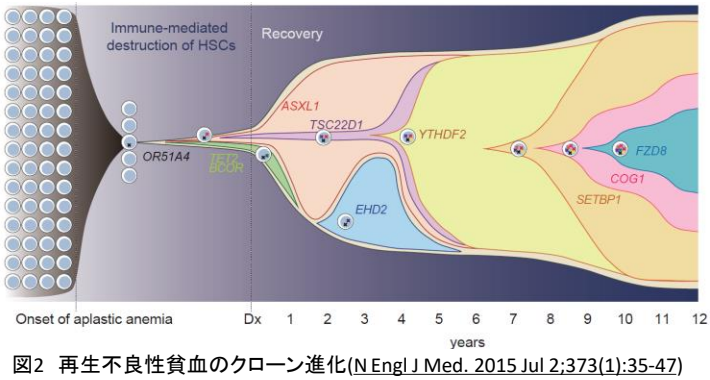


図2　再生不良性貧血のクローン進化(N Engl J Med. 2015 Jul 2;373(1):35-47)

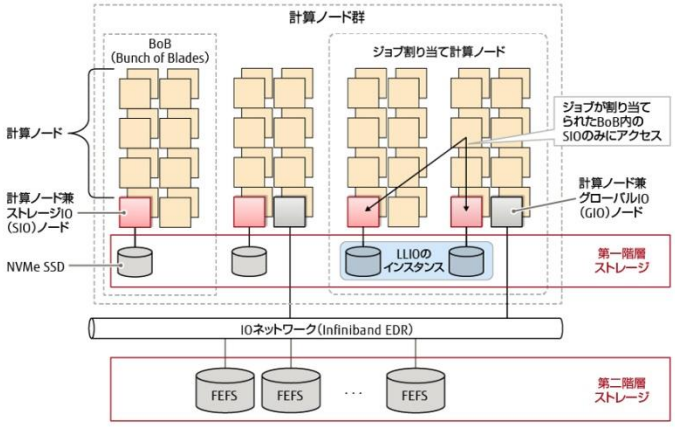


図3　スーパーコンピュータ「富岳」の階層化ファイルシステム
(https://www.fujitsu.com/jp/about/resources/publications/technicalreview/2020-03/article05.html)

Acknowledgements

本研究は、文部科学省「富岳」成果創出加速プログラム「大規模データ解析と人工知能技術によるがんの起源と多様性の解明」（課題番号：hp200138）の一環として実施されたものです。また、本研究の一部は、（スーパーコンピュータ「京」／スーパーコンピュータ「富岳」）の計算資源の提供を受け、実施しました。