ポスト「京」重点課題 2

個別化・予防医療を支援する統合計算生命科学

NEWS LETTER

Vol.14

Contents

INTEGRATED

COMPUTATIONAL

LIFE

SCIENCE

TO SUPPORT

PERSONALIZED AND

PREVENTIVE MEDICINE

ポスト「京」
重点課題2
個別化・予防医療を支援する統合計算生命科学

■ Research Report

Subtheme A

次世代のがん治療に向けた変異検出の高精度化

サブ課題 A
東京大学医科学研究所　　森山 卓也
（ポスト「京」重点課題 2 サブ課題 A 協力者）

がんはゲノムの変異により生じる疾患で，がん組織の中にはゲノムの変異が異なる多様な細胞集団を持つ（多様性を持つ）ことが知られています。通常、分子標的薬による治療においては、特定の変異を持つがん細胞を標的に治療を行いますが、多様な細胞集団の中の一部が標的にされずに残ることにより、治療は困難となります。
上記の問題を解決するには、次世代シーケンサーから読み取ったゲノムの情報（シーケンスデータ）から一部の細胞集団を捉える必要があります。そのためには、低いアレル頻度の変異を捉える高精度な変異検出手法が必要であり、この手法の開発はがんゲノムにおける最重要課題です。
我々は、サブ課題 A「大量シーケンスによるがんの個性と時間的・空間的多様性・起源の解明」の中で、高精度な変異検出手法の開発を行ってきました。今回はその内容や取り組みに関して簡単に報告します。

シーケンスデータの特性に基づいた変異検出手法

変異検出に最も良く利用される、Illumina のシーケンサーでは、以下の手順でゲノム情報が読み取られます(図1)。
(a) 組織からDNA分子（約30億塩基対）を抽出する。
(b) 抽出したDNA分子（約30億塩基対）をランダムに200から500塩基対程度に断片化する。
(c) 断片化したDNAの両端から100塩基対程度のDNA配列（ペアエンドリード）を読み取る。

(b) の過程において断片化した DNA 分子のサイズが小さい場合、ペアエンドリードに重複が生じ、同じ箇所の DNA 配列が二回にわたって読み取られます。 我々は、この特性が変異検出に応用可能であることを、全エキソンシーケンスデータ [1] を用いて検証し(図 2)、変異検出手法 OVarCall [2] を開発しました。

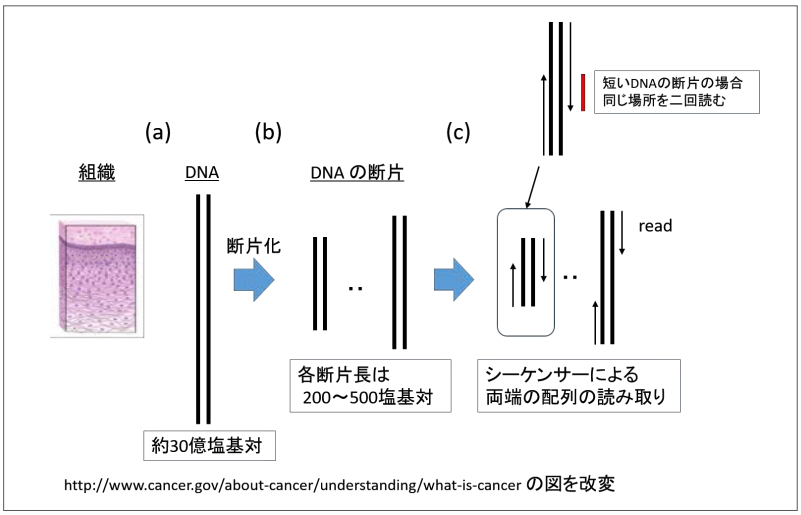

図1： Illumina シーケンサーのワークフロー

複数の特性が利用可能な統計モデリングによる変異検出手法

上記のペアエンドリードの重複以外にも、変異検出の高精度化に役に立つ特性が知られていて、全ゲノムシーケンスデータからの変異検出に有効であることが知られていました [3]。しかし上記 OVarCall においてはその特性まで加味できてはおらず、それら複数の特性をもとに変異検出を行う手法は存在しませんでした。

そこで、さらなる高精度化に向け、複数の特性を加味可能なベイズ統計的な枠組みを構築し、これを元に変異検出手法 OHVarfinDer [4] を開発しました。性能評価においては、The Cancer Genome

■ Research Report

Subtheme A

Atlas が公開しているデータセット [5] も利用し、ほとんどの場合で既存手法に比べて高い性能を示すことを確認しました。

がん組織のシーケンスデータを複数箇所利用する手法

がんは、様々な変異を獲得しつつ増大し、進化していきます。この変異を蓄積する過程は系統樹という木構造を用いて表現されます [6]。図 3 の系統樹の各節点はある時刻における細胞集団、各節点間を結ぶ線分（枝）は蓄積した変異に対応します。ヒトのゲノムは約 30 億塩基対と非常に大きいものであるため、一つの体細胞変異は系統樹の枝のどれか一つにしか対応しないと、広く想定されています (Infinite site assumption) [7]。この想定のもとでは、各変異に対してどの細胞集団が変異を持つかを表す変異パターンに対して制約を導くことができます。この制約は、変異検出の高精度化を行う上で、非常に使い勝手の良い特性になることが期待できます。

しかし、上記の特性を利用するには、一人の患者さんから、がん組織のシーケンスデータを複数箇所利用する必要があります。複数のがん組織のシーケンスデータを想定していない OVarCall や OHVarfinDer では、その特性まで利用が不可能です。そこで我々は、複数箇所のシーケンスデータを用い、上記の特性を使いつつ変異検出を高精度化する手法の開発を現在進行形で行っています。

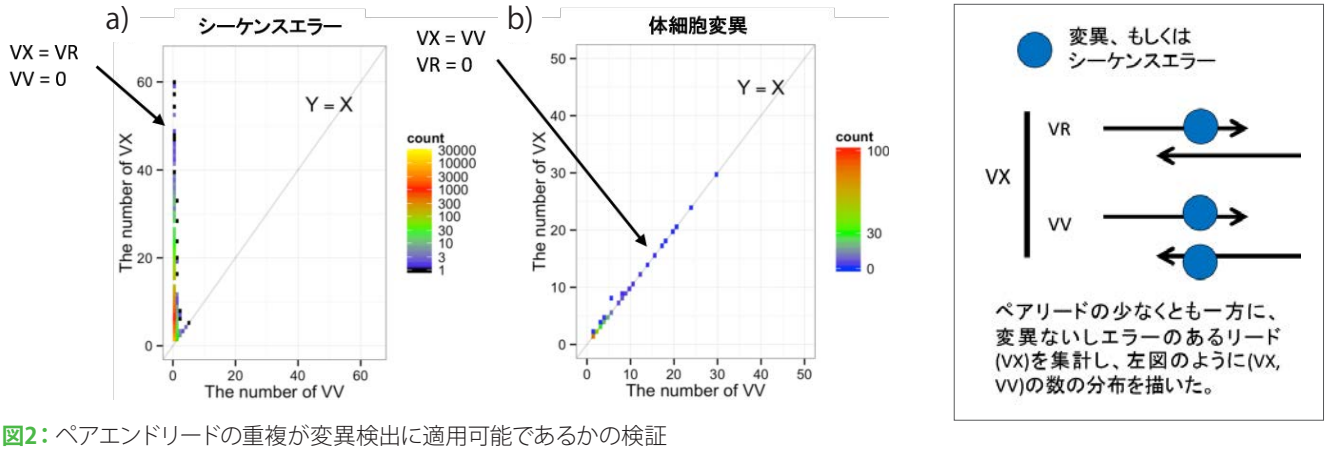

図2： ペアエンドリードの重複が変異検出に適用可能であるかの検証
実データから、変異もしくはシークエンスエラーを持つ重複リードを集め、特性を調べた。
a) シークエンスエラーが起きている箇所であれば、ほとんどの場合、リードの片方にエラーが生じる。
b) 真の体細胞変異が存在する箇所ではほとんど全てのリードで両方に変異が観測される。

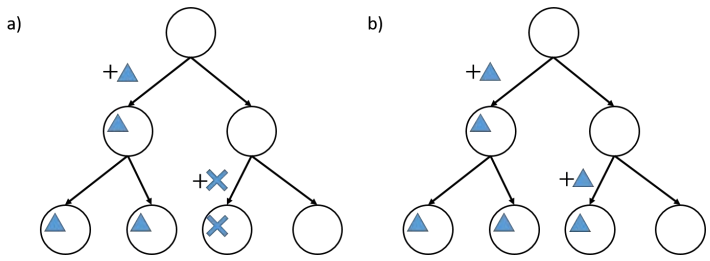

図3： 系統樹とInfinite site assumption
節点が細胞集団、節点同士を結ぶ線分が蓄積した変異に対応する。a) Infinite site assumption を満たす場合の系統樹。各変異はせいぜい、一つの線分にのみ対応する。この場合、変異がどの細胞に存在するかのパターンが、系統樹上の線分の数以下に抑えられるという制約が導ける。b) Infinite site assumption を満たさない場合の系統樹。ある変異に対しては一つ以上の線分に対応する。

＜参考文献＞
[1] Yoshida, K. et al. (2011) Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*, 478, 64–69.
[2] Moriyama, T. et al. (2017) OVarCall: Bayesian mutation calling method utilizing overlapping paired-end reads. *IEEE Transactions on NanoBioscience*, 16(2), 116-122.
[3] Usuyama, N. et al. (2014) HapMuC: somatic mutation calling using heterozygous germ line variants near candidate mutations. *Bioinformatics* (Oxford, England), 30(23), 3302–9.
[4] Moriyama, T. et al. (2019) A Bayesian model integration for mutation calling through data partitioning. *Bioinformatics* (Oxford, England)にて論文採択済み
[5] https://gdc.cancer.gov/resources-tcga-users/tcga-mutation-calling-benchmark-4-files
[6] Gusfield, D. (1991) Efficient algorithms for inferring evolutionary trees. *Networks*, 21, 19-28.
[7] Kimura, M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61, 893–903

第56回日本臨床分子医学会学術集会

- 日程：4月26日(金) ～ 27日(土)
- 場所：ポートメッセなごや・交流センター（愛知県名古屋市）
- Web：http://www.mtoyou.jp/jsmm56/

○ 特別講演1

- 日時：4月26日(金) 11:10 ～ 12:10
- 演者：小川 誠司 (京都大学大学院医学研究科)
- タイトル：悪性腫瘍発生の本質に迫るアプローチ(仮)

第30回 日本医学会総会 2019 中部 学術集会

- 日程：4月27日(土) ～ 29日(月・祝)
- 場所：名古屋国際会議場、ウインクあいちほか(愛知県名古屋市)
- Web：http://isoukai2019.jp/

○ 講演

最先端ヒトゲノム情報の医療への応用(柱1-2-1)

- 日時：4月27日(土) 16:10 ～ 18:10
- 場所：第6会場 名古屋国際会議場 白鳥ホール(南)
- 演者：小川 誠司 (京都大学大学院医学研究科)

市民公開講座
～ AIが命を救い、新薬を開発する！～人工知能が切り開く未来医療(柱1-1-4)

- 日程：4月28日(日) 14:00 ～ 16:00
- 場所：第28会場 ウインクあいち 大ホール
- 演者：宮野 悟 (東京大学医科学研究所)
- タイトル：人工知能パワースーツを装着した医師によるがんゲノム医療(予定)

第62回日本腎臓学会学術総会

- 日程：6月21日(金) ～ 23日(日)
- 場所：名古屋国際会議場(愛知県名古屋市)
- Web：http://jsn62.umin.jp/

総会長主導企画2「腎臓内科学へのAI・ICT技術の応用」

○ 講演

- 日時：6月22日(土) 15:40 ～ 17:40
- 場所：第1会場(名古屋国際会議場 1号館 2 階)
- 演者：宮野 悟 (東京大学医科学研究所)

第27回日本乳癌学会学術総会

乳がん患者の心と身体のケア
―乳がんゲノム医療と支持医療―

- 日程：7月11日(木) ～ 13日(土)
- 場所：京王プラザホテル・新宿NSビル（東京都新宿区）
- Web：http://www.congre.co.jp/jbcs2019/

○ 特別講演

- 演者：小川 誠司 (京都大学大学院医学研究科)

○ シンポジウム

- 名称：Augmented Intelligence (拡張知能)
- 日時：2019年7月11日(木) 8:30 ～ 10:30
- 会場：第5会場（京王プラザホテル4階 錦）
- 演者：宮野 悟 (東京大学医科学研究所)

PICK UP! Research Report からの用語解説

アレル頻度
集団内における対立遺伝子 (アレル) の割合。仮に集団内で Aa:AA = 1:1 の場合、a のアレル頻度は、1/4 = 25% になる。

ベイズ統計
ベイズ主義に基づく統計学の考え方。頻度主義では母数は固定されているものとするが、ベイズ主義では母数に対して確率分布を設定する。

文部科学省　ポスト「京」開発事業
重点的に取り組むべき社会的・科学的課題に関するアプリケーション開発・研究開発
重点課題2　個別化・予防医療を支援する統合計算生命科学
Integrated Computational Life Science to Support Personalized and Preventive Medicine

■ 問い合わせ先
国立大学法人東京大学医科学研究所　ヒトゲノム解析センター　DNA 情報解析分野
ポスト「京」重点課題 2 個別化・予防医療を支援する統合計算生命科学 事務局
〒 108-8639 東京都港区白金台 4-6-1　TEL：03-5449-5615　FAX：03-5449-5442
E-mail：icls-office@hgc.jp　　URL：http://postk.hgc.jp/

ポスト「京」重点課題は、国家基盤技術としてスーパーコンピュータ「京」の後継機となるポスト「京」を活用し、国家的に解決を目指す社会的・科学的課題に戦略的に取り組み、世界を先導する成果の創出を目指す文部科学省の事業です。重点課題 2「個別化・予防医療を支援する統合計算生命科学」は、東京大学医科学研究所を代表機関として、ポスト「京」によって初めて実現できる「情報の技術」、「物理の原理の応用」、そして「ビッグデータの活用」により、病態の理解と効果的な治療の探索法の研究を行い、その成果を個別化・予防医療へ返す支援基盤となる統合計算生命科学を確立することを目的としています。

COVER PHOTOGRAPH : Wild flowers
NEWS LETTER VOL.14 2019.3