（平成 27 年度研究報告書）

26-A-6　がん情報生物学・生物統計学研究基盤の構築

柴田　龍弘
国立がん研究センター　研究所

研究の分類・属性

発がん・がん生物学分野

研究の概要

本研究班は、次世代型解読技術の進歩に伴うがんオミックスビッグデータの集積とその活用による新たながん医療を支える情報解析基盤並びに基礎研究分野の確立を目指し、がんゲノム解析等のオミックスデータを精度高く分析する方法論の調査・検討・開発と、シュミレーションや確率統計学モデルを駆使してがんの特性である臨床的・生物学的多様性を数理学的に解析する新たな基礎研究分野に関する研究を実証的に行い、がん研究分野における情報解析支援基盤の構築と国際競争力強化を進める。また若手研究者を中心としたネットワークを作り、がん情報解析分野における人材育成を計る。

平成 27 年度研究経費

15,999 千円

研究班の組織

研究者名	所属研究機関名・職名	分担研究課題名
柴田　龍弘	国立がん研究センター・研究所・がんゲノミクス研究分野・分野長	研究の統括・高速シークエンサーを用いた情報解析の検証
十時　泰	国立がん研究センター・研究所・がんゲノミクス研究分野・ユニット長	がんオミックスビッグデータの高精度な分析・解釈・知識抽出のための手法に関する研究・開発
鈴木　穣	東京大学・大学院・新領域創成化学研究科・教授	がんオミックスビッグデータの高精度な分析・解釈・知識抽出のための手法に関する研究・開発
三浦 史仁	九州大学・医学研究院・講師	がんオミックスビッグデータの高精度な分析・解釈・知識抽出のための手法に関する研究・開発
辻　真吾	東京大学・先端科学技術研究センター・特任助教	がんオミックスビッグデータの高精度な分析・解釈・知識抽出のための手法に関する研究・開発

濱　奈津子	国立がん研究センター・早期探索臨床研究センター・研究員	がんオミックスビッグデータの高精度な分析・解釈・知識抽出のための手法に関する研究・開発
波江野　洋	九州大学・大学院理学研究院生物科学部門・助教	数理学的・統計学的解析（モデル構築、シュミレーション解析など）のための基礎的検討並びに解析手法の研究・開発
井上　雅世	産業技術総合研究所・創薬分子プロファイリング研究センター・研究員	数理学的・統計学的解析（モデル構築、シュミレーション解析など）のための基礎的検討並びに解析手法の研究・開発

研究の目的と到達目標及び実績要点

全期間

（目的と到達目標）

【研究の背景】第２世代あるいは第３世代シークエンス技術の進歩により、がんゲノム・エピゲノム情報の多量化・重層化が進んでいる。今後国内外の主要ながん臨床研究拠点では、こうした大量のがんオミックスビッグデータの集積とそれを活用した新たな先駆的がん医療が進められる事が予想され、当センターにおいても高度情報化医療に対応した体制構築のために必須な情報解析基盤の確立が急務である。更に近年免疫療法に関する臨床知見が増加し、腫瘍免疫環境を評価する情報解析ツールの必要性も高まっている。他方、シークエンス解読コストの減少に伴い、がんの特性である臨床的・生物学的多様性の分子背景となるがんゲノム・エピゲノムの時空間的変遷（発がん過程あるいは転移・再発・治療抵抗性獲得等におけるがんの分子的進化）や宿主微小環境を含めた腫瘍内多様性を数理学的に解析するためのデータ取得が１細胞解析を含め実現可能となり、がんの数論学的解析に向けたモデル構築、あるいはシュミレーション・確率統計学的手法を駆使した予測医学が、がん医療の更なる発展を支える新しい基礎研究フロンティアとして期待されている。

【本研究班の目的と到達目標】センター内外の専門家による研究と議論によって、がんゲノム解析等のオミックス解析ビッグデータを精度高く分析する方法の開発と、がんの様々な特性に関する新たな知見や、臨床への展開について、数理生物学の専門家の視点から検討するためのバイオインフォマティクス基盤の確立を目指す。

第２年次

（到達目標）

１．ゲノム・エピゲノムといったがんオミックスビッグデータの高精度な分析・解釈・知識抽出のための手法を研究・開発し、当センターの情報解析基盤の確立を目指す。腫瘍免疫環境を評価するような解析ツールの評価を行う。更にナノポアシークエンスを含めた次世代シークエンスデータ解析について検討を継続し、また１細胞解析については本センター内の情報解析への活用を試みる。

２．がんの特性の数論的理解や予測医療を含めた臨床応用に向けた数理学的・統計学的解析（モデル構築、シュミレーション解析など）のための基礎的検討並びに解析手法を研究・開発し、実際の臨床検体によるパラメータ調整や解釈について検討を進め、当センターにおける基礎研究への活用を含めた実証的な検討も進める。

（当該年評価時点の実績要点）

1. 新たな手法を導入しつつ、センター内における解析（オミックス解析・pathway 解析）10 件（475 検体）における情報解析の支援を行った。
2. 全ゲノムバイサルファイトシークエンシング（Whole-genome bisulfite sequencing、WGBS）によって得られ

た単塩基解像度かつ高精度なDNAメチロームデータから、注目する組織や細胞に特異的なメチル化状態を示すゲノム領域（Differentially Methylated Region）の同定を効率良く進める情報処理技術の開発を行った。

3. ナノポアシークエンサーを用いたゲノム変異を検出するプログラムの作成を行った。肺腺がん培養細胞株由来のmRNAを用い、EGFR、KRAS等の本がん種において代表的なドライバー遺伝子についてその全転写産物領域のナノポアシークエンスを行い、その結果について評価を行った。
4. PacBio RSを用いて全ゲノム解読データについて、short readによるエラー修正を行い、構造異常に関する解析手法の開発を進めた。
5. 近年注目を集めるDeep Learningを用いて、サンプルラベルの有無に関わらず、大規模なマルチオミックスデータから、新たな知識を抽出する手法について検討を行い、Deep Learningが同一のがん種のサブタイプ解析などに利用出来る可能性が得られた。
6. 腫瘍免疫環境解析ツールの評価・開発に着手し、まずT細胞受容体レパトア解析について既存のツールを調査・検討した。PCR biasを除去でき、より正確なコピー数判定を行うことができるMIGEC toolを用いて、臨床検体を使った性能検証を行った。
7. がん進展ダイナミクスの数理学的解析による微小転移巣の予測を検討した。原発腫瘍が増殖していく過程で確率的に転移を起こす数理モデルを解析することによって、診断時の原発腫瘍の大きさ、増殖率、転移巣の増殖率、遺伝子変異確率、転移率に依存して、診断時における転移巣のサイズ毎の数の期待値を求める理論式を導いた
8. 遺伝子をその発現量制御方法に注目し、合成過程による量制御が重要なクラス、分解過程による量制御が重要なクラス、翻訳後修飾が重要なクラスに分類する手法開発を進め、多くのがん関連遺伝子が分解過程による量制御が重要なクラスに分類されていることを明らかにした。

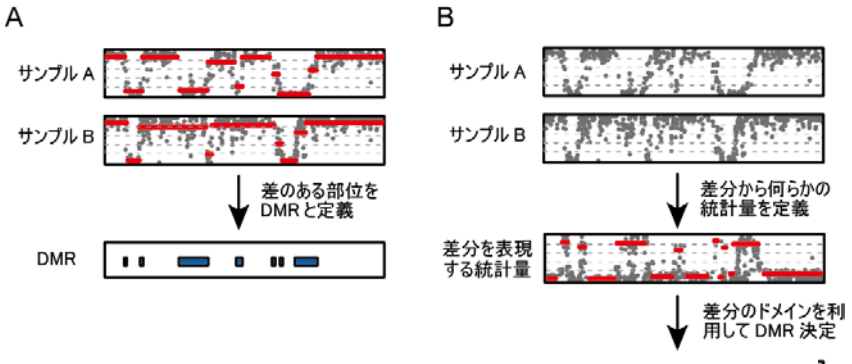
研究成果と考察

第2年次評価時点

1-1. エピゲノムデータの解析手法の開発

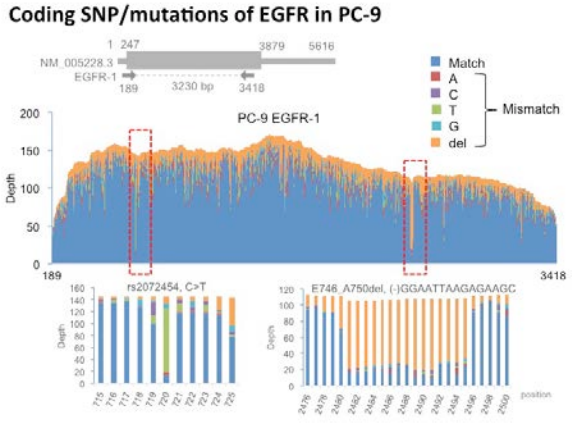
(1)メチル化率に変化が生じた領域（Differentially Methylated Region：DMR）を同定する手法開発

WGBSによるメチロームデータが増大するにつれ、注目する組織や細胞に特異的なメチル化状態を示すゲノム領域（DMR）の同定を効率良く進める情報処理技術が必要となっている。変曲点検出がDMRの効率的検出にも応用可能かどうかについて検討した。まず、それぞれのサンプルの平均メチル化率のデータから変化点検出に基づいてドメインを検出した後に、サンプル間でドメインの境界を比較し、サンプル間で境界のズレが確認された場合にその領域をDMRと定義する方法（図1A）と、各シトシンのメチル化率をサンプル間で比較し、その比較から得られた個々のシトシンのパラメータを変曲点検出に適用してドメインを定義する方法（図1B）の二つを考えた。前者は境界線の揺らぎをどの程度まで許すのか、欠測データに対する処理の仕方など、最適化や考慮すべきパラメータの数が多いことがわかった。一方、後者は、境界線の問題は変化点検出アルゴリズムにより解決され、また欠測に対してはそれを考慮したパラメータの計算法を考えることが可能であるため、より容易にDMRの同定プログラムを開発することが可能であると判断された。そこで、後者を実現するためのメチル化率の変化を表現するパラメータの定義に関して検討を進めた。



1-2. ナノポアシークエンサー技術の検討

ナノポアシークエンサーを用いたゲノム変異を検出するプログラムの作成を行った。肺腺がん培養細胞株由来のmRNAを用い、EGFR、KRAS等の本がん種において代表的なド



ライバー遺伝子についてその全転写産物領域のシークエンスを行った。PCR 増幅産物混合物のシークエンスから 2195 本の配列を取得した。これらのうち 1253 本はアラインメントプログラム LAST を用いて、200 以上のマッピングスコアで参照ゲノム配列にマッピングすることが可能であった。個々の配列について、その塩基配列読み取り精度を検証したところ、平均 QV 値が 9.5、参照ゲノムとの一致率は平均 83%であった。個々には精度が十分でないこれらの配列ではあるが、ほとんどの標的領域塩基について 50 本以上の配列でカバーされていた。これらについてそのコンセンサス配列を生成し、そのコンセンサス配列を参照ゲノム配列と比較したところ、100%の検出感度（1 か所の SNP と 1 か所の 15 塩基欠失を検出；一方でその他の塩基については擬陽性なし）で、EGFR 中のコーディング領域に存在するゲノム変異を検出することが可能であった（右図参照）。これらの変異はサンガーシークエンス、イルミナシークエンスで確認された。一方で UTR 領域については、5 か所、擬陽性の変異が検出された。これらの周辺配列を観察したところ、ホモポリマー中に存在する 1 塩基欠失が誤って検出されていることが分かった。現在、他の遺伝子も含めて、さらに正確に変異の検出が可能なパラメーターの最適化を進めている。

1-3 Single molecule sequencer による解析手法の開発

PacBio RS を用いて全ゲノム解読データについて、short read によるエラー修正を行い、構造異常に関する解析手法の開発を進めた。

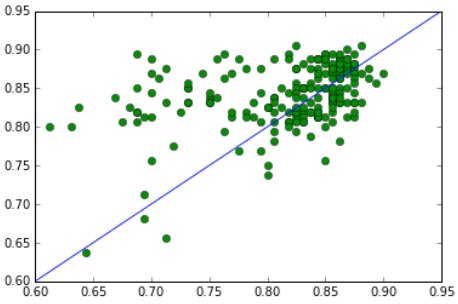
1-4 機械学習アルゴリズムを用いたがんゲノムデータからの知識抽出手法開発

近年注目を集める Deep Learning を用いて、サンプルラベルの有無に関わらず、大規模なマルチオミックスデータから、新たな知見を得ることができる可能性を示した。肝臓がんのマルチオミックスデータ（体細胞変異、遺伝子発現、DNA メチル化）を用いて、早期肝癌と進行肝癌の判定について検討を行った。

教師有り・無しの場合ともに、学習させた Deep Learning モデルに、サンプルのデータを入力として与え、出力されたサンプルの数値データを考察することで、モデルからどのような知見が得られるかを検討した。その結果、教師有り学習の場合は、サンプルラベルに対する良い予測モデルとなっていることが確認できた上に、進行肝癌の中にサブグループがあることを示唆する結果となり、これは TP53 と CTNNB1 の体細胞変異の数の違いとして裏付けられた。一方、教師無し学習の場合は、早期と進行のラベルを用いずに学習モデルを作ったが、完成したモデルが、早期肝癌と進行肝癌の予測モデルとしても利用出来ることが分かり、与えていない情報に Deep Learning アルゴリズムが気付いたとも言える結果となった。

教師なし学習の場合

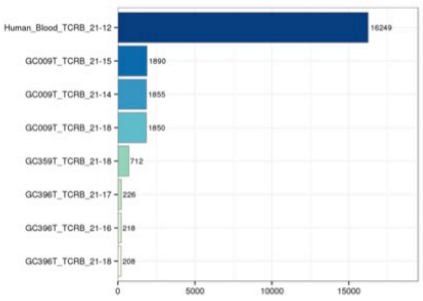
Deep Learning の手法の 1 つである、autoencoder を使って、自分自身のデータを再現するモデルを構築した。遺伝子発現と DNA メチル化のデータをランダムに選び出すことによって、200 個の仮想的なデータセットを作り、出来上がったモデルを通して数値を変換したデータ群を K-means クラスタリングで 2 分し、早期癌と進行癌のラベルとどれくらい合致するかを計算した。右図は、横軸に変換前のデータ、縦軸に変換後のデータをとって、200 データセットごとの予測率をプロットしたものである。

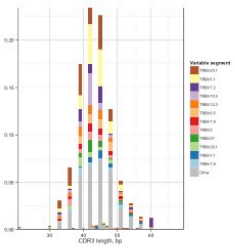


こうした結果は Deep Learning が同一のがん種のサブタイプ解析などに利用出来る可能性を示唆している。今後、どのような因子が、モデルの構築に重要な役割を果たしているのかを調査することによって、大規模マルチオミックスデータ解析のための基盤を、さらに強固なもの出来ると考えている。

1-5 腫瘍免疫環境解析ツールの評価

T 細胞受容体において抗原が結合する領域は CDR3 と呼ばれ、塩基の挿入および欠失がランダムに起こり、多様性を獲得している。通常血中に存在するリンパ球と比較して、腫瘍浸潤リンパ球は、腫瘍抗原を認識してクローナルに増殖しており、ユニークな CDR3 配列を持った T リンパ球の頻度分布（レパトア）を正確に評価する必要がある。これまでシークエンス技術を使った T 細胞レパトア解析ツールは複数報告されているが、





Count	%	V	J	D	Total reads
42	0.08	TRBV13	TRBJ2-2	TRBD1	56,400
30	0.06	TRBV9	TRBJ1-1	TRBD1	17,273
19	0.03	TRBV5-1	TRBJ2-5	TRBD1	17,928
17	0.03	TRBV20-1	TRBJ2-3	TRBD2	13,409
12	0.02	TRBV3-1	TRBJ2-1	TRBD1, TRBD2	8,730

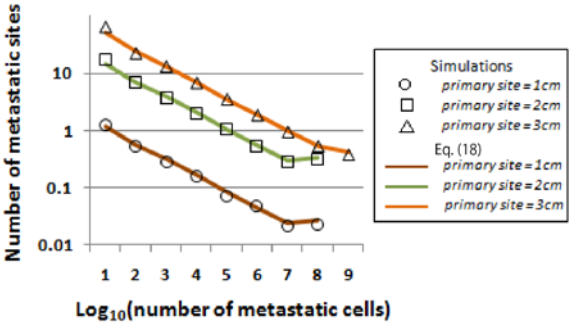
文献調査の結果、PCR bias を除去でき、より正確なコピー数判定を行うことができる MIGEC tool を用いて、臨床検体を使った性能検証を行った。ヒト正常末梢血並びにがん検体から RNA を抽出後、各 RNA 分子に tag を付け、その後 RTPCR で CDR3 領域を増幅することで、バーコード化したライブラリーを作成した。このバーコードを使うことによって、PCR duplication を除去し、正確な頻度推定が可能になる。右図に結果を示すが、正常末梢血に比較して腫瘍検体では全体のクローン数の減少を認め、また特定の TCR 配列を持ったクローンが濃縮していることが確認できた。今後本解析パイプラインについて、他の手法との比較検証を行うと同時に、センター内の解析支援にも使えるような自動化を進める。

1-6 センターにおける情報解析支援

全ゲノム・エクソーム解読, RNAseq, methylome 解析など 10 件（475 検体）の解析について情報解析支援を行った。

2-1 がん進化に関する数理学的解析

数理学的視点から、原発腫瘍が増殖していく過程で確率的に転移を起こす数理モデルを解析することによって、理論上転移が存在する確率や転移巣の数の期待値を求めた。その結果、診断時の原発腫瘍の大きさ、増殖率、転移巣の増殖率、遺伝子変異確率、転移率に依存して、診断時における転移巣のサイズ毎の数の期待値を求める理論式を導いた（図）。



左図 理論式と確率シミュレーションによる転移巣のサイズ毎の数の期待値に関する分布の予測
それぞれの色は診断時の腫瘍サイズを表し、直線は理論式による予測、点は確率シミュレーションによる結果を表す。診断時における腫瘍サイズがどのような場合においても、サイズが大きい転移巣の数はサイズが小さい転移巣の数よりも少なくなる傾向を示す。

サイズの大きい転移巣ほど転移巣の数が少ないという結果が得られた。定性的な示唆として、現在の腫瘍検出機器によって見つからないサイズの転移巣が存在する確率は、検出出来るサイズの転移巣が存在する確率よりも大きくなっており、転移が検出されない場合でも、微小な転移巣が存在した場合への対応を考慮する必要があると言える。
また、過去の研究で推定された膵臓癌の定量的パラメータを用いて、既報の膵癌診断時に転移を持たない症例群と転移巣を持つ症例群の生存曲線を再現した（下図）。

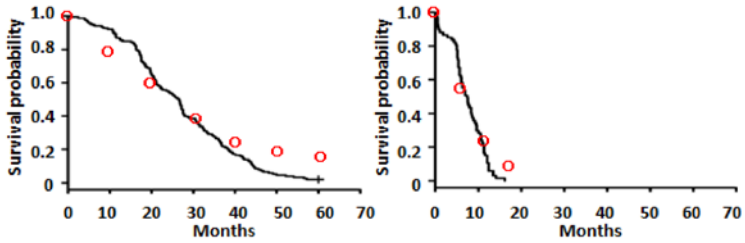
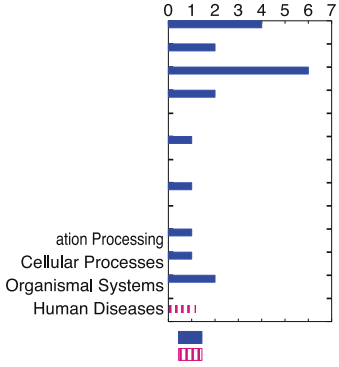


図 膵癌症例の生存曲線の再現
赤い点が膵癌症例の生存曲線に関する臨床データを表す。黒い曲線が理論式による生存曲線を表す。左図は診断時に転移がないと判断され、根治手術を行った症例群（Oshima et al., Ann Surg. 2013)。右図は診断時に転移が存在しており手術を行わなかった症例群（Cunningham et al., J. Clin. Oncol. 2009)。
これらの結果は論文として Scientific Reports 誌に発表した。(Yamamoto, et al. Haeno. 2015, Sci. Rep.)。

2-2 オミックス間相互関係の解析

遺伝子をその発現量制御方法に注目し、合成過程による量制御が重要なクラス、分解過程による量制御が重要なクラス、翻訳後修飾が重要なクラスに分類する手法の開発をおこなった。各クラスの特徴抽出として Paythway 解析をおこなったところ、クラス毎に異なる特徴的遺伝子機能と関連していること(右図参照)、多くのがん関連遺伝子が分解過程による量制御が重要なクラスに分類されていることが示された。



倫理面への配慮

解析方法の検証研究に使用する解析データの取得に関しては、「ヒトゲノム・遺伝子解析研究に関する倫理指針」並びに「疫学研究に関する倫理指針」を遵守し、個人情報の保護等、試料等提供者の人権擁護上の配慮を十分に行い、また患者にとって不利益や危険性が起こらないように留意し、検体の使用について同意が得られた検体のみを用いる。検体については、国立がん研究センター中央病院並びに東病院個人情報管理者によって連結可能匿名化の作業を行った後に、解析を行う。解析データの取得に関する研究については、国立がん研究センター中央病院並びに東病院の倫理審査委員会の承認を得る。

本研究に関連する、本研究期間中の主な発表論文等

第２年次

（雑誌論文）

国立がん研究センター研究開発費による成果であることが記載されているもの

1. Nakamura H, Arai Y, Totoki Y, Shirota T, Elzawahry A, Kato M, Hama N, Hosoda F, Urushidate T, Ohashi S, Hiraoka N, Ojima H, Shimada K, Okusaka T, Kosuge T, Miyagawa S, Shibata T. Genomic spectra of biliary tract cancer. **Nat Genet**, 2015, 47:1003-10.
2. Kataoka K, Nagata Y, Kitanaka A, Shiraishi Y, Shimamura T, Yasunaga JI, Totoki Y, Chiba K, Sato-Otsubo A, Nagae G, Ishii R, Muto S, Kotani S, Watatani Y, Takeda J, Sanada M, Tanaka H, Suzuki H, Sato Y, Shiozawa Y, Yoshizato T, Yoshida K, Makishima H, Iwanaga M, Ma G, Nosaka K, Hishizawa M, Itonaga H, Imaizumi Y, Munakata W, Ogasawara H, Sato T, Sasai K, Muramoto K, Penova M, Kawaguchi T, Nakamura H, Hama N, Shide K, Kubuki Y, Hidaka T, Kameda T, Nakamaki T, Ishiyama K, Miyawaki S, Yoon SS, Tobinai K, Miyazaki Y, Takaori-Kondo A, Matsuda F, Takeuchi K, Nureki O, Aburatani H, Watanabe T, Shibata T, Matsuoka M, Miyano S, Shimoda K, Ogawa S. Integrated molecular analysis of adult T cell leukemia/lymphoma. **Nat Genet**. 2015, 47:1304-15.
3. Yokoyama T, Miura F, Araki H, Okamura K, Ito T. Changepoint detection in base-resolution methylome data reveals a robust signature of methylated domain landscape. **BMC genomics**. 2015;16(1):594. 査読有り
4. Inoue M, Horimoto, K. Gene function determines how the level of gene expression is regulated. (投稿中)

国立がん研究センター研究開発費による成果であることが記載はないが、関連するもの

1. Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, Mustonen V, Gonzalez-Perez A, Pearson J, Sander C, Raphael1 BJ, Marks1 DS, Ouellette BF, Valencia A, Bader GD, Boutros PC, Stuart JM, Linding R, Lopez-Bigas N, Stein LD. Pathway and network analysis of cancer genomes. **Nat Methods**, 2015 12:615-21.
2. Iwakawa R, Kohno T, Totoki Y, Shibata T, Tsuchihara K, Mimaki S, Tsuta K, Narita Y, Nishikawa R, Noguchi M, Harris CC, Robels AI, Yamaguchi R, Imoto S, Miyano S, Totsuka H, Yoshida T, Yokota J. Expression and clinical significance of genes frequently mutated in small cell lung cancers defined by whole exome/RNA sequencing. **Carcinogenesis**, 2015, 36:616-21.
3. Yamamoto KN, Nakamura A, Haeno H. The evolution of tumor metastasis during clonal expansion with alterations in metastasis driver genes. **Sci. Rep**. 2015 30,5: 15886.