様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ５ 月 ２４ 日現在

機関番号：１２６０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００３９６

研究課題名（和文）網羅的塩基配列解読データを用いたコンタミネーションの検出と影響解析手法の開発

研究課題名（英文）Developing a computational method for microbial contaminant detection and functional inference using next-generation sequencing data

研究代表者

朴　聖俊（Park, Sung-Joon）

東京大学・医科学研究所・特任講師

研究者番号：４０７５９４１１

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：多種多様な細胞実験が繰り返し行われている現代生物学において、目的細胞や関連サンプルが細菌・ウイルスにさらされる危険性は一層増しており、感染・汚染の防止と検出は極めて重要な課題である。本研究では、目的細胞の次世代シークエンシング（NGS）データに含まれている外来性ゲノム、すなわちコンタミを網羅的かつ高確度で検出するアルゴリズムを開発・公開した。これにより、例えば、遺伝子発現とコンタミが一度にプロファイリングでき、その関連性の深化した解析が可能となった。

研究成果の学術的意義や社会的意義
開発した手法の公開とその解析結果のデータベース化を行ったことで、既存研究データから見られるコンタミの種類とその混入度分布がわかり、ホスト細胞に与える外来性ゲノムのインパクト推定が容易となった。これは、細胞培養方法と実験試薬の見直しや既存研究データの再解釈に資するものであり、さらに、核酸増幅法などの既存検査方法の代替方法として、例えば、再生医療製品の安全性と品質管理、新規細菌・ウイルス同定にも応用可能であることから、今後の応用発展を図りたい。

研究成果の概要（英文）：In modern biology, cells are routinely manipulated by various experimental techniques under a range of conditions. These increase the risk of exposure of the cells to microorganisms that cause unexpected molecular changes and misunderstanding. Therefore, the prevention and detection of microbial contamination is a critical issue. In this study, we developed a computational method to comprehensively and accurately detect contaminants that present in the next-generation sequencing (NGS) data. This method can profile, for example, transcriptome and contamination simultaneously. Therefore, this method allows us to perform deeper analyses of the interaction with contaminants.

研究分野：バイオインフォマティクス

キーワード：コンタミネーション　バイオインフォマティクス　次世代シークエンシング

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

本研究開始当初、次世代シークエンサー（NGS）が様々な分野で欠かせない中核基盤技術として確立され、誰でも手軽く自動的に迅速なゲノム解析が行われつつあったことは論を待たない。一方では、シークエンシング手法、実験材料などが多種多様化し、元サンプルと抽出データの品質管理に学術的疑義が生じていた。例えば、1000 人ゲノムプロジェクトの全ゲノム解読データの約 7%（PMID:24872843）、NCBI GEO 登録のトランスクリプトームデータの約 11%（PMID: 25712092）にマイコプラズマ感染の疑いがある、といった報告もなされていた。しかし、研究サンプル・データの品質を体系的に評価する計算アルゴリズムがなく、単に NGS で読まれた配列データ（リード）を様々な参照ゲノム配列に繰り返しマッピングするアプローチに留まっていた。したがって、外来性ゲノム（コンタミ）の詳細なプロファイリングと混入原因究明、また、ホスト細胞への影響の解析に資する情報基盤の開発が急を要した。

２．研究の目的

一般的に、参照ゲノムにアライメントできない NGS リードは実験ノイズとして解析から除外されるが、このようなリードを詳細に解析することで標的細胞の細菌・ウイルス感染・汚染状況を評価することが可能である。本研究では、NGS データに存在する外来性ゲノムの網羅的かつ高確度にプロファイリングする手法と、感染・汚染起因の異常発現遺伝子を推定する手法を開発する。そして、開発手法と既存データの解析結果を一般に広く提供する情報基盤を構築し、実験細胞の汚染リスク評価と改善につなげることが目的である。

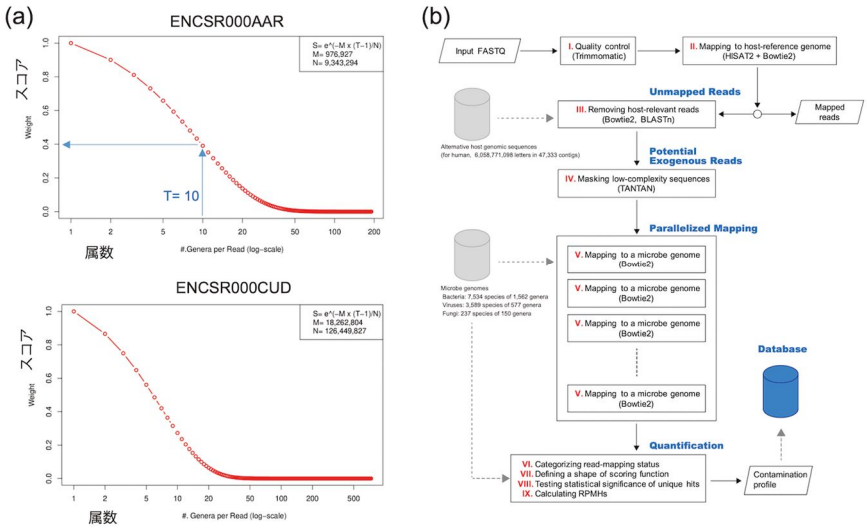

図１：(a)複数属にマップされたリードにペナルティを与える例。(b)本研究で構築したパイプライン

３．研究の方法

（１）動的スコア関数によるコンタミの定量化

まず、細菌・ウイルス由来のコンタミリードが既存アラインメントツールでどれほど正確に検出できるかを検証した。ここでは、バクテリア、ウイルス、真菌の完全ゲノム（2,289 属の 11,360 種）を NCBI RefSeq から用意して微生物ゲノム DB を作成し、ここから人工 NGS リードセットをランダムサンプリングで生成した（1,000 reads x 10 species /set）。このランダムリードセットが各種アラインメントツール（BLASTn, Bowtie, Bowtie2, BWA, NovoAlign）の様々なパラメータで元ゲノムに戻る割合を計算する、いわば「Reversion Test」を行った。

その結果、17%以上の種はそれぞれの 5%以上のリードが誤ったゲノムに戻ることが分かった。つまり、微生物種間の高い配列類似度を考慮した解析手法が必要である。そこで、独自の定量スコア関数を設計した。提案スコア関数では、リード１個が１属の微生物ゲノムにマップされたとき（ユニークヒット）は 1.0 とカウントするが、多属のゲノムへ重複してマップされた場合（マルチヒット）は属数に応じて指数的にペナルティを与えている。指数関数の傾斜は、解析対象の NGS データによって異なるように設計し、ユニークヒットの多いケースではマルチヒットのペナルティを強くして抑えるが、ユニークヒットの少ないケースではマルチヒットのリード数をより多く取り入れるためにペナルティを弱くする（図 1a）。

（２）検出コンタミの有意性評価

例えば、解析データのトータルリードの内、N 個が外来性ゲノム由来リードで、その内 X 個が微生物種 Y のユニークヒットである場合、この「X 個」というのはどれほど意味のある数値であるのだろうか？適当に作成した人工リードでも Y のユニークヒット X 個が見つかるとしても、感

染種 *Y*は信頼できるコンタミだろうか？
　ここでは、*Y*を除いた微生物 DB からランダムに用意した *N* 個の人工リードを *Y*ゲノムにマッピングし、ユニークヒットを数える。この操作を繰り返すことで *Y*との高い配列類似度に由来する偶然なユニークヒットの Z 分布を推定し、実データから観測した *X*個の統計的有意性を評価した。

（3）解析パイプラインの構築
上述のスコアリングと統計処理を取り入れた解析パイプラインを高性能スパコン上に構築した（図 1b）。本パイプラインでは既存手法と同じく、まず、対象リードのホストゲノムへのマッピングを繰り返してコンタミ候補リードを絞る。その後、コンタミ候補リードを微生物 DB を用いてマッピングと統計処理を行って有意なコンタミ種と属をリストする。Single-end RNA-seq 1 千万リードの場合だと 5 時間ほどで終了する。

（4）機械学習によるホスト-コンタミ相関解析
トランスクリプトーム（RNA-seq）の場合、本パイプラインは遺伝子発現プロファイルとコンタミプロファイルを同時に出力する。この二つのプロファイルを関連付けることでコンタミのインパクトが推定できる。ここでは、非負値行列因子分解（Non-negative Matrix Factorization, NMF）の拡張版である Joint NMF を適用した。Join NMF は、コンタミ量の増減と転写量増減を同時にクラスタリングすることで両者の関係性を推定する方法である。本研究では、RNA-seq のレプリケーションに繰り返し Joint NMF とネットワーク解析を行って、レプリケーションで共通して発現上昇する遺伝子を抽出し、それらの機能解析を行った。

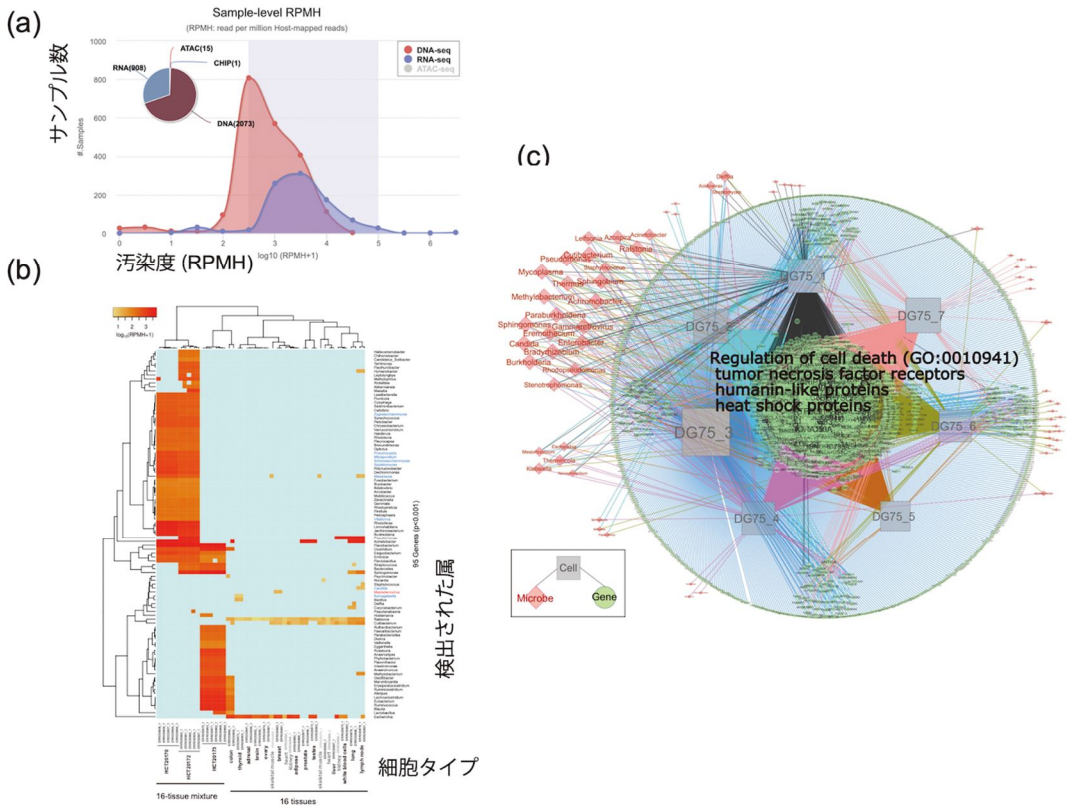

図 2：(a)DNA-seq と RNA-seq における RPMH 値の分布。(b)高度な細胞操作による汚染リスクの増大（Illumina BodyMap2.0 RNA-seq 例、右の大きいクラスタが 16 ヒト組織、左が 16-tissue mixture）。(c)Joint NMF とネットワーク解析による異常発現遺伝子の同定（マイコプラズマ感染 DG-75 細胞株 7 つのレプリケーションの例）。

4．研究成果
（1）コンタミ量の推定
まず、NCBI の SRA、1000 人ゲノムプロジェクトと GEUVADIS プロジェクトなどから 3500 サンプル以上をダウンロードして解析を行った。その結果、RNA-seq の場合は 10^3–10^4 RPMH（Reads Per Million Host-mapped read）、つまり、ホストゲノムが百万個読まれると 1,000 から 10,000 個のコンタミリードが含まれることがわかった。この値は全ゲノムデータでもさほど変わらなかった（図 2a）。

（2）検出されたコンタミの傾向

次に大量公共データにみられるコンタミの特徴と傾向を考察し、次のことがわかった。①ホスト由来のリード数が減るとコンタミ由来のリードが増える。つまり、NGS 実験失敗とコンタミとの関連性が推察された。②細胞培養、分化誘導などの操作を繰り返すとコンタミの種類と強度が増した（図 2b）。③最も頻繁に検出される細菌に Escherichia、Cutibacterium, Pseudomonas などのラボ環境由来菌が多く含まれていた。④5%のサンプルから統計的に有意なマイコプラズマ汚染が見つかった。⑤Illumina 社フラットフォームのスパイクインである PhiX174 は予想よりはるか多いサンプルに顕著に残っていて、さらに、PhiX174 由来のリードは G4、Alpha3 Microvirus とマルチヒットすることが分かった。つまり、この種のゲノム解析に PhiX174 スパイクインは不向きであることがわかった。

（３）コンタミに応答するホスト遺伝子の同定

最後に、マイコプラズマに感染した間葉系幹細胞の遺伝子発現解析を行ったところ、マイコプラズマ感染に特異的に応答する遺伝子は ER-associated degradation (ERAD) pathway に関わることを明らかにした。しかし、細胞種が変わるとこのようなシグネチャーも変化することがわかり、Joint NMF とネットワーク解析を駆使してシステマティックな応答遺伝子の同定を行った。その結果、マイコプラズマ感染 DG-75 細胞株の場合、感染細胞は炎症系とアポトーシス抑制因子の過剰発現の状態にあることがわかった（図 2c）。すなわち、当該細胞が細胞死から逃れ、正常に回収されてシークエンシングされていたのは、こういったホスト側の遺伝子応答によるものだったのかもしれない。

（４）考察

本研究で開発した手法と解析結果は OpenContami(https://openlooper.hgc.jp/opencontami/) に公開している。以上の研究成果から、当初の目的であった「コンタミネーションの検出と影響解析の手法開発」は達成されたといえる。しかし、コンタミ原因の特定には至っていない。これは、予期しない微生物の混入経路と時期が多種多様であり、さらに常在菌も含んでいる為で、極めて難題である。例えば、リードクオリティを見積もるために PhiX174 をスパイクインするように、実験細胞をいれない“ブランク”シークエンシングも並行することで、ある程度の特定は可能であると考えられる。また、実験条件などの詳細な情報をデータベース化して管理することも必要であり、本研究で開発した情報基盤がその実現に活用されるであろう。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Sung-Joon Park, Satoru Onizuka, Masahide Seki, Yutaka Suzuki, Takanori Iwata, and Kenta Nakai	４．巻 17
２．論文標題 A systematic sequencing-based approach for microbial contaminant detection and functional inference	５．発行年 2019年
３．雑誌名 BMC Biology	６．最初と最後の頁 72
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1186/s12915-019-0690-0	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Sung-Joon Park	４．巻 3
２．論文標題 Genome-Wide Scanning of Gene Expression	５．発行年 2019年
３．雑誌名 Encyclopedia of Bioinformatics and Computational Biology	６．最初と最後の頁 452-462
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/B978-0-12-809633-8.20132-5	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計7件（うち招待講演　1件／うち国際学会　4件）

１．発表者名 鬼塚 理, 朴 聖俊, 貝淵 信之, 妻沼 有香, 安藤 智博, 中井 謙太, 岩田 隆紀
２．発表標題 歯根膜組織由来間葉系幹細胞シートによる歯周組織再生
３．学会等名 第19回日本再生医療学会総会
４．発表年 2020年

１．発表者名 Sung-Joon Park, Satoru Onizuka, Masahide Seki, Yutaka Suzuki, Takanori Iwata and Kenta Nakai
２．発表標題 OpenContami: A Web-based Application for Detecting Microbial Contaminants in Next-generation Sequencing
３．学会等名 第8回生命医薬情報学連合大会
４．発表年 2019年

１．発表者名
Sung-Joon Park and Kenta Nakai

２．発表標題
Host-unmapped NGS reads shape the prevalence of microbial contamination

３．学会等名
The Biology of Genomes（国際学会）

４．発表年
2019年

１．発表者名
Sung-Joon Park and Kenta Nakai

２．発表標題
Profiling microbial contamination by exploiting host-unmapped NGS reads

３．学会等名
Revolutionizing Next-Generation Sequencing 2019 (RNGS19)（国際学会）

４．発表年
2019年

１．発表者名
Sung-Joon Park and Kenta Nakai

２．発表標題
Applying Joint Non-Negative Matrix Factorization to Functional Analysis of Microbial Infection

３．学会等名
The 17th International Conference On BioInformatics（国際学会）

４．発表年
2018年

１．発表者名
Sung-Joon Park

２．発表標題
Profiling of Cell Microbial Contamination by NGS data

３．学会等名
配列解析シンポジウム、日本バイオインフォマティクス学会（招待講演）

４．発表年
2018年

１．発表者名
Sung-Joon Park

２．発表標題
Development of a Pipeline for Contamination Profiling of Cells with Next Generation Sequencing Data

３．学会等名
The 28th International Conference on Genome Informatics Workshop (GIW)/BIOINFO 2017（国際学会）

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
OpenContami
https://openlooper.hgc.jp/opencontami/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考