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　 ６ 月 １９ 日現在

機関番号：８２５０４
研究種目：若手研究(B)
研究期間：2017～2019
課題番号：１７Ｋ１５０４７

研究課題名（和文）機械学習によるピロールイミダゾールポリアミドでゲノム結合の解析

研究課題名（英文）Analyzing Genomic Binding of Pyrrole-Imidazole Polyamides by Machine Learning

研究代表者
ＬＩＮ　ＪＡＳＯＮ（Lin, Jason）

千葉県がんセンター（研究所）・がん治療開発グループ　がん遺伝創薬研究室・研究員

研究者番号：８０７７４１２４

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：癌遺伝子の主蛋白質を標的にすることでがん治療法が開発されているが、治療薬の開発が難しい標的分子も多い。ピロールイミダゾールポリアミド(PIP)は、副溝に結合してDNAに直接標的する化合物であり、様々な癌遺伝子を阻害することが確認された。しかし、大半のPIPは8~10塩基認識であり、ゲノム内には多くの結合配列が存在し、PIPの治療効果、毒性や副作用などの安全性を予測することが困難と考えられる。本研究では、バイオインフォマティクス及び機械学習法により次世代シーケンシング及び発現プロファイリングデータ解析を基づいて、PIPにおけるがんゲノムとの結合の評価及び臨床的な副作用の予測モデルを行った。

研究成果の学術的意義や社会的意義
がん治療では、標的タンパク質構造によって、大半のがん標的に直接阻害することが難しい。ピロールイミダゾールポリアミド(PIP)は、DNA塩基認識によってDNA副溝結合を経由し、様々な癌遺伝子を阻害することが確認された。塩基認識では数学的なリミットがあり、ゲノム内には多くの結合配列が存在する。臨床実用へのため、PIPの治療効果、毒性や副作用などの安全性を予測しなければならない。本研究では、人工知能における機械学習法により、がんゲノムデータを解析し、PIPにおけるがんゲノムとの結合の評価及び臨床的な副作用の予測モデルを行った。本研究の結果によりPIPががん治療薬剤として実現することを期待される。

研究成果の概要（英文）：The high sequence specificity of minor groove-binding N-methylpyrrole-N-methylimidazole polyamides have made significant advances in cancer and disease biology, yet there have been few comprehensive reports on their off-target effects, most likely as a consequence of the lack of available tools in evaluating genomic binding, an essential aspect that has gone seriously underexplored. Compared to other N-heterocycles, the off-target effects of these polyamides and their specificity for the DNA minor groove and primary base pair recognition require the development of new analytical methods, which are missing in the field today. This project seeks to incorporate methods in computational biology and machine learning in the analysis of next-generation sequencing and expression profiling data to decipher off-target effects of these polyamides.

研究分野：ゲノミクス・エピゲノミクス

キーワード：ケミカルバイオロジー　バイオインフォマティクス　ゲノミクス　エピゲノミクス

１．研究開始当初の背景

革新的ながん治療法が開発視されているが、「標的困難」な癌分子も多く存在する。我々は、この問題に遺伝子産物でなく、遺伝子配列を標的にすることで解決を試みている。一方、既存の核酸を標的とする RNA 干渉や CRISPR/Cas などのゲノム編集技術では、不安定性や必要となる遺伝子工学の繁雑さ等の問題点を抱える。我々は siRNA よりも安定であり、CRISPR/Cas のような遺伝子工学技術が不要な配列特異的 DNA 副溝結合化合物 (minor groove binder)、ピロールイミダゾールポリアミド(PIP)を独自に臨床応用に向けて開発している。PIP は A/T と C/G 塩基に対する配列特異的結合により活性阻害が困難ながん遺伝子の発現を抑制でき、我々は既に KRAS の G12D/V 変異を認識する PIP(KR12)を合成し、ヒト大腸癌細胞株においてイン・ビトロ及びイン・ビボで増殖を阻害すること[1]や、ヒストン脱アセチル化酵素阻害薬の SAHA を付加した PIP がエピゲノム修飾を変更できることを実証している。しかし、大半の PIP は 8-10bp 認識であり、ゲノム内には多くの結合配列が存在するため、生物学的活性を詳細に評価し、実際の結合領域を予測することが困難であり、PIP が薬剤候補として臨床試験がまだ開始されない。

２．研究の目的

PIP は、創薬困難な標的分子に対しても標的遺伝子配列特異的(8–10bp)に DNA 副溝に結合することで標的化できる化合物である。我々はゲノム内での PIP 結合配列を同定するため PIP を用いた ChIP-seq 次世代シーケンシング法(NGS)を新規に開発し、KR12 が KRAS 変異大腸癌細胞において G12D/V 変異に結合し細胞死を誘導したことから、我々は KR12 のゲノム結合部位の推定法の開発(無架橋結合 Chem-seq シークエンシング)を行った[2]。しかし、従来のピーク判断手法を用いた解析では NGS 配列リードのピーク数を過小評価してしまうことが明らかとなった。一方、PIP を用いた Chem-seq ではクロスリンクを行わないため、配列リードのサイズやピークの位置がバラバラなヘテロな集団となっていることが予想され、従来の SPC ではピーク数の過小評価につながるとも考えられた。これらの問題点を解決するために、我々は 300-1200bp の様々なスライディングウィンドウを設定し、パラメータフリーのコルモゴロフ-スミルノフ検定により統計学的解析を行った。実際、この方法によりピーク数の過小評価は軽減したが、現在のデザインによる大規模データセット処理に対応出来なかった。さらに、コルモゴロフ-スミルノフ検定は二つの母集団の確率分布が異なるものであるかどうかについての検出力は優れているが、Chem-seq データの母集団分布が複雑なため、検出力が低下した可能性がある。本課題では、PIP を用いた Chem-seq という NGS 手法を使って、機械学習法を基づいてゲノム結合サイトを判断するアルゴリズムやゲノム発現プロファイルに及ぼすフェノタイプ影響を予測する新規解析プラットフォームを構築することを目的とする。

３．研究の方法

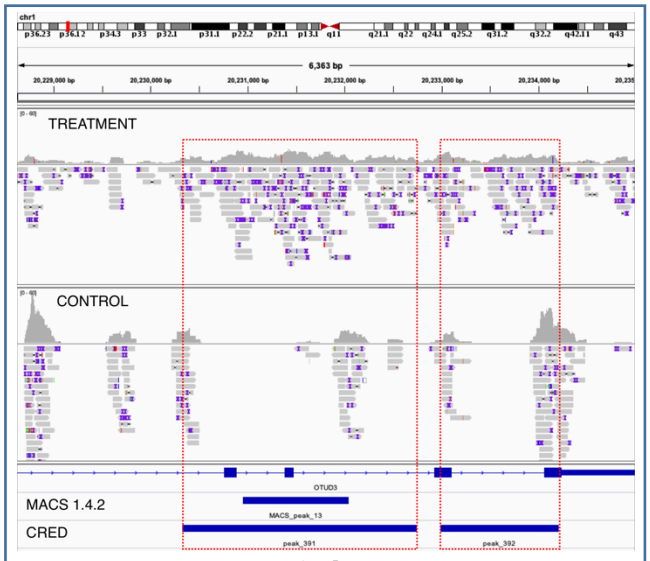

図 1、CRED によるピーク判定。「TREATMENT」という部位は、PIP を用いた沈殿した DNA 断片であり、「CONTROL」は PIP を利用しない DNA 断片。CRED は、既存のピーク判定ソフトである MACS より多くの特定部位を認識できる（文献６より）。

本研究では、PIP 化合物の正確な結合部位同定を目的に Chem-seq 実験データから高精度にピーク推定を行うアルゴリズムを開発し、発現プロファイリングによる臨床的な副作用などのフェノタイプ影響を予測する新規解析プラットフォームを構築することを目指す。本アルゴリズムでは、二階段の解析ステップとなる。ピーク測定速度を上げるために、統計学的な分布モデルを基づいて、最初に「CRED」というソフトウェアを作成し、ゲノムエンリッチメント領域を測定して、次に「rfPIPeak」というソフトウェアを開発し、機械学習法によるピークを詳しく調べて、PIP でのゲノム結合サイトの解析を試みる。従来の手法は、分析ウィンドウ 1 つごとに、分布パラメータを推定して、ピークを検出するが、パフォーマンスを向上させ可能な C 言語で作成した CRED では、スライディングウィンドウの 1200bp を確立し、その設定の中で結果としてより多く結合部位を検出できた(図 1)。PIP を用いた Chem-seq の特徴として、アデニンの 1 塩基対のみをアルキル化し、DNA とのアダクトを形成しプルダウンできる点にある。このような特徴下では、ChIP-seq 解析するために最適化した MACS は、PIP を用いた Chem-seq データに分析優位性が失われるが、本方法では MACS よりコントールピークの検出数が高くなると考えられる。さらに、既存の Chem-seq データより、高品質を維持するフィルタリングで繰り返し配列リードを階層化し、配列リードを取得し、機械

学習用のトレーニングデータセットを構築する。配列リードデータはカテゴリ変数へ変更し、トレーニングセットに割り当てられた分類に基づいて、実験データを分類する。リードによる PIP 結合(及び非結合)場所を手動割当して、ピーク（結合サイト）またはノンピーク（結合しない領域）を集計し、機械学習用の特徴的な塩基長、塩基長ウィンドウ内の配列リード頻度、極大と極小値などの統計情報を収集する。処理されたデータより教師あり機械学習法であるランダムフォレストを基づいてプログラミング言語の R と Perl で実装される。SVM は計算生物学における様々なアプリケーションに用いられる教師あり機械学習法であるが、ノイズのある高容量のデータセットで、カバレッジパフォーマンスの高い ChIP-seq のデータでは、予測能力が低下する傾向が認められる。また、Chem-seq 実験で PIP と結合する配列パターンでは不均一なパイルアップパターン（同一配列を含むフラグメントの集積）が見られるため、ランダムフォレストが最適な分類・予測アルゴリズムではないかと考えられる。アルゴリズムの選択は産総研人工知能研究センターの指導のもと検討し行った。

ちなみに発現マイクロアレイの情報を用いて、KEGG [4]パスウェイデータベースを基づいて PIP 投与による発現プロファイリングへの影響を各パスウェイが単位として、可能なオフターゲット（OT）遺伝子が同定でき(図 2)、その結果によって臨床的な副作用などのフェノタイプ変化の予測が可能と考えられる。薬剤副作用データベースや GEO データセットより抽出した薬剤候補マイクロアレイ結果を統合し、データクリーニングを行い、発現変化による副作用を影響するトレーニングデータセットを作成した。その後、機械学習法であるランダムフォレスト法による副作用の予測するアルゴリズムを確立し、複数の PIP で発生した共同の副作用を予測した。

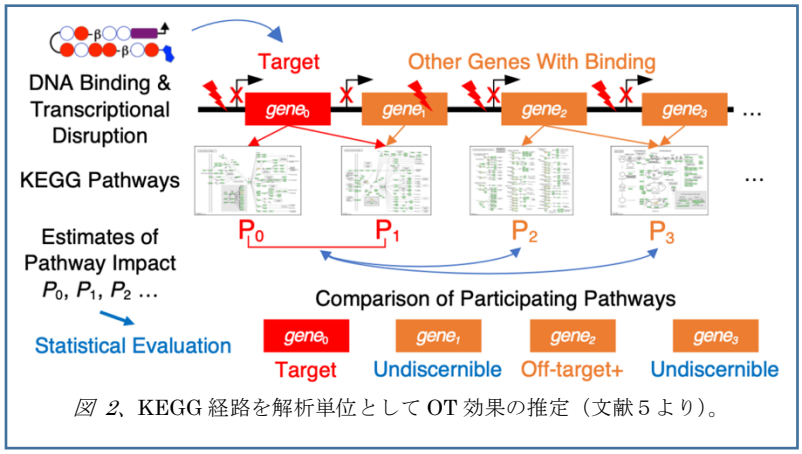

図 2、KEGG 経路を解析単位として OT 効果の推定（文献５より）。

さらにマウス実験を用いて検証実験を行い、予測された副作用における肝機能マーカーである AST・ALT 変化及び食欲不振（食量の減少）も検証した。現在、同定アルゴリズムの検証実験は続いて行う。複数のアルキル化剤 CBI を付加した PIP にさらにビオチンを付加し、ヒト人工染色体(HAC)における DNA 断片のエンリッチメントを行い、NGS でフラグメント解読データを取得、HAC 配列に再構築し、結合配列を判定する。ヒト癌遺伝子である *KRAS* を含む HAC を選択し、CRED や rfPIPeak の精度及びパーフォマンスを評価している。ちなみに開発されたアルゴリズムのソースコードをリリース後、主要なコンポーネントを最適化している（C 言語での書き直す）。

４．研究成果

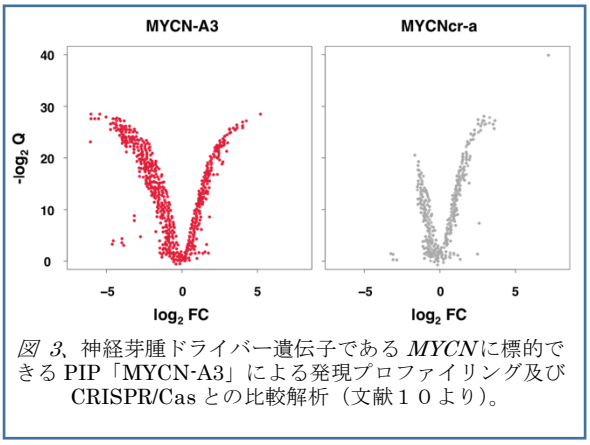

図 3、神経芽腫ドライバー遺伝子である *MYCN* に標的できる PIP「MYCN-A3」による発現プロファイリング及び CRISPR/Cas との比較解析（文献１０より）。

本課程から、PIP を用いて「Chem-seq」という NGS 手法(図 1D)の開発を進め、2016 年に最初のデザインに基づく機械学習法を用いた Chem-seq ピーク推定手法を開発、2019 年に発現プロファイリングをパスウェイが単位として解析し、オフターゲット(OT)マーカーが予測でき、さらに機械学習法を利用して PIP ゲノム阻害が臨床的な副作用などのフェノタイプに及ぼす影響の測定法を発表した[5]。CRED という Chem-seq サイト検出について段階的なアルゴリズムを JOSS 誌に掲載された[6]。ランダムフォレスト法を用いてエンリッチメント領域に結合サイトの推定は、一部（既存の CRAN パッケージ kmeRs に基づいて「kmeRs2」という配列モチーフ類似性行列ツール及び「rfPIPeak」というランダムフォレストによるピーク精製同定ツール）が GitHub で公開され[7,8]、またヒト人工染色体での検証実験を終わり次第論文を投稿する予定である。OT マーカーの推測手法は、プログラミング言語 R のパッケージ「pipoft」に統合され、現在オープンソースの形で GitHub に公開されている[9]。

本課程で開発した発現アレイ解析手法を用いて、MYCN に標的する PIP である「MYCN-A3」が神経芽腫細胞株における発現プロファイリングに及ぼす影響(図 3)が Cancer Research 刊に掲載された[10]。本研究の成果も複数の国内や海外の国際学会（日本癌学会、国際計算生物学会）で

発表され、さらに 2020 年３月に PIP でのゲノム結合によるオフターゲット、副作用及びゲノム解析手法について蓄積した知見(図 4)がレビューとして「Biomolecules」誌に掲載された[11]。今まで PIP の分野では、NGS 法による PIP でのオフターゲット可能性の検討は、Chem-seq 解析プラットホーム改善が可能と期待できる。

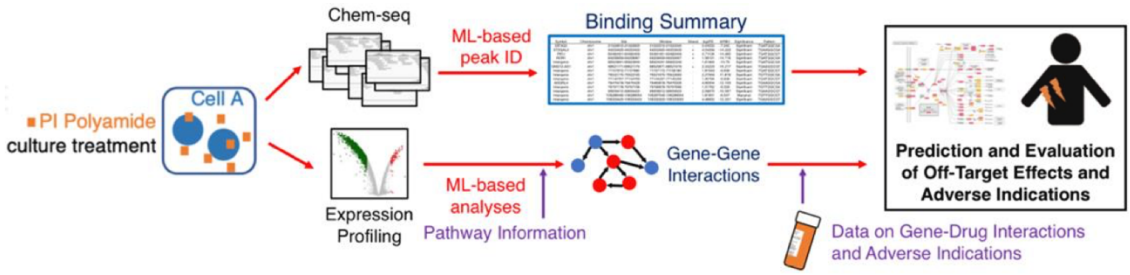

図 4、機械学習法による PIP における OT 効果や可能な臨床的副作用の予測（文献１１より）。

<引用文献>

[1] Hiraoka K et al, Inhibition of KRAS codon 12 mutants using a novel DNA-alkylating pyrrole-imidazole polyamide conjugate, *Nat. Commun.*, 6, 2015, 6706.
[2] Lin J et al, Identification of Binding Targets of a Pyrrole-Imidazole Polyamide KR12 in the LS180 Colorectal Cancer Genome, *PLoS ONE*, 11, 2016, e0165581.
[3] Zhang Y et al, Model-based Analysis of ChIP-seq (MACS), *Genome Biol.*, 9, 2008, R137.
[4] Kanehisa M, Goto S, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 28, 2000, 27-30.
[5] Lin J et al, Estimating genome-wide off-target effects for pyrrole-imidazole polyamide binding by a pathway-based expression profiling approach, *PLoS ONE*, 14, 2019, e0215247.
[6] Lin J et al, CRED: a rapid peak caller for Chem-seq data, *J. Open Source Softw.*, 4, 2019, 1423.
[7] *kmeRs2*: k-Mers Similarity Score Matrix. (github.com/jlincbio/kmeRs2)
[8] *rfPIPeak*: R package for Chem-seq peak calling by random forest. (github.com/jlincbio/rfpipeak)
[9] *pipoft*: R package for estimating effects of off-target binding by pyrrole-imidazole polyamides. (github.com/jlincbio/pipoft)
[10] Yoda H et al, Direct Targeting of MYCN Gene Amplification by Site-Specific DNA Alkylation in Neuroblastoma, *Cancer Res.*, 79, 2019, 830-840.
[11] Lin J, Nagase H. The Road Not Taken With Pyrrole-Imidazole Polyamides: Off-Target Effects and Genomic Binding, *Biomolecules*, 10, 2020, 544.

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　4件／うち国際共著　3件／うちオープンアクセス　3件）

１．著者名 Lin Jason、Kuo Tony、Horton Paul、Nagase Hiroki	４．巻 4
２．論文標題 CRED: a rapid peak caller for Chem-seq data	５．発行年 2019年
３．雑誌名 Journal of Open Source Software	６．最初と最後の頁 1423
掲載論文のDOI（デジタルオブジェクト識別子） 10.21105/joss.01423	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 該当する

１．著者名 Lin Jason、Krishnamurthy Sakthisri、Yoda Hiroyuki、Shinozaki Yoshinao、Watanabe Takayoshi、Koshikawa Nobuko、Takatori Atsushi、Horton Paul、Nagase Hiroki	４．巻 14
２．論文標題 Estimating genome-wide off-target effects for pyrrole-imidazole polyamide binding by a pathway-based expression profiling approach	５．発行年 2019年
３．雑誌名 PLOS ONE	６．最初と最後の頁 e0215247
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pone.0215247	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 該当する

１．著者名 Yoda Hiroyuki、Inoue Takahiro、Shinozaki Yoshinao、Lin Jason、Watanabe Takayoshi、Koshikawa Nobuko、Takatori Atsushi、Nagase Hiroki	４．巻 79
２．論文標題 Direct Targeting of MYCN Gene Amplification by Site-Specific DNA Alkylation in Neuroblastoma	５．発行年 2018年
３．雑誌名 Cancer Research	６．最初と最後の頁 830～840
掲載論文のDOI（デジタルオブジェクト識別子） 10.1158/0008-5472.CAN-18-1198	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 該当する

１．著者名 Inoue Takahiro、Shimozato Osamu、Matsuo Nina、Mori Yusuke、Shinozaki Yoshinao、Lin Jason、Watanabe Takayoshi、Takatori Atsushi、Koshikawa Nobuko、Ozaki Toshinori、Nagase Hiroki	４．巻 26
２．論文標題 Hydrophobic structure of hairpin ten-ring pyrrole-imidazole polyamides enhances tumor tissue accumulation/retention in vivo	５．発行年 2018年
３．雑誌名 Bioorganic & Medicinal Chemistry	６．最初と最後の頁 2337～2344
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.bmc.2018.03.029	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 －

1．著者名 Lin Jason、Nagase Hiroki	4．巻 10
2．論文標題 The Road Not Taken with Pyrrole-Imidazole Polyamides: Off-Target Effects and Genomic Binding	5．発行年 2020年
3．雑誌名 Biomolecules	6．最初と最後の頁 544～544
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/biom10040544	査読の有無 無
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11件（うち招待講演　0件／うち国際学会　6件）

1．発表者名 Lin Jason, Takatori Atsushi, Horton Paul, Nagase Hiroki
2．発表標題 Chem-seq: Evaluation of Genomewide Binding Effects of DNA Minor Groove-Binding Pyrrole Imidazole Polyamides
3．学会等名 RECOMB/ISCB Conference on Regulatory & Systems Genomics with DREAM Challenges（国際学会）
4．発表年 2018年

1．発表者名 Lin Jason，平岡桐子，Krishnamurthy Sakthisri，井上貴博，養田裕行，篠崎喜脩，渡部隆義，高取敦志，越川信子，永瀬浩喜
2．発表標題 ピロールイミダゾールポリアミドによるがん治療の評価：Chem-seq
3．学会等名 第77回日本癌学会学術総会（国際学会）
4．発表年 2018年

1．発表者名 Lin Jason，平岡桐子，養田裕行，Krishnamurthy Sakthisri，篠崎喜脩，渡部隆義，高取敦志，越川信子，Horton Paul，永瀬浩喜
2．発表標題 ゲノム解析によるピロールイミダゾールポリアミドで生化学的機能および臨床的副作用の予測
3．学会等名 第27回日本癌病態治療研究会
4．発表年 2018年

１．発表者名
Lin Jason，奥村和弘，寺島裕也，遠田悦子，板倉明司，松島綱治，永瀬浩喜

２．発表標題
臨床検体を用いた網羅的遺伝子発現解析によるがん増悪化分子フロントに関連するバイオマーカーの開発研究

３．学会等名
AMEDがん若手研究者ワークショップ

４．発表年
2019年

１．発表者名
J. Lin, A. Tataktori, K. Hiraoka, H. Yoda, S. Krishnamurthy, T. Inoue, T. Watanabe, T. Kuo, Y. Shinozaki, N. Koshikawa, P. Horton, H. Nagase

２．発表標題
Design of a next-generation affinity-enrichment Chem-seq sequencing procedure to assess the biochemistry of minor-groove-binding pyrrole-imidazole polyamides

３．学会等名
2017年度生命科学系学会合同年次大会（国際学会）

４．発表年
2017年

１．発表者名
J. Lin, K. Hiraoka, H. Yoda, A. Takatori, T. Watanabe, T. Kuo, Y. Shinozaki, N. Koshikawa, P. Horton, H. Nagase

２．発表標題
Assessing oncotherapeutic efficacy of minor-groove-binding pyrrole-imidazole polyamides by Chem-seq sequencing

３．学会等名
"第76回日本癌学会学術総会（国際学会）

４．発表年
2017年

１．発表者名
Jason Lin

２．発表標題
機械学習によるピロールイミダゾールポリアミドでのゲノム結合の解析

３．学会等名
"文部科学省先端モデル動物支援プラットフォーム・若手支援技術講習会

４．発表年
2017年

１．発表者名
Jason Lin, Sakthisri Krishnamurthy, Hiroyuki Yoda, Yoshinao Shinozaki, Takayoshi Watanabe, Nobuko Koshikawa, Atsushi Takatori, Paul Horton, Hiroki Nagase

２．発表標題
発現プロファイリングによるDNA副溝結合剤のピロールイミダゾールポリアミドでの オフターゲット及び副作用の評価・予測

３．学会等名
第78回日本癌学会学術集会（国際学会）

４．発表年
2019年

１．発表者名
Jason Lin, Atsushi Takatori, Hiroki Nagase, Paul Horton

２．発表標題
Estimation of Phenotypes and Side Effects from DNA Minor-Groove-Binding Pyrrole-Imidazole Polyamides

３．学会等名
27th Conference on Intelligent Systems for Molecular Biology and the 18th European Conference on Computational Biology (ISMB/ECCB 2019)（国際学会）

４．発表年
2019年

１．発表者名
リン・ジェイソン、クリシュナムーティ・サクテイシリ、養田裕行、篠崎喜脩、渡部隆義、越川信子、高取敦志、ホートン・ポール、永瀬浩喜

２．発表標題
ピロールイミダゾールポリアミドでの非特異的なゲノム結合および副作用の予測

３．学会等名
第28回日本癌病態治療研究会

４．発表年
2019年

１．発表者名
リン・ジェイソン、渡部隆義、杉山弘、永瀬浩喜

２．発表標題
塩基配列選択的な新規エピゲノム制御分子PIP-SAHAにおける機械学習法による遺伝子発現量影響の予測

３．学会等名
第5回AMEDがん若手研究者ワークショップ

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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