

令和 5 年 6 月 8 日現在

機関番号：12102

研究種目：若手研究

研究期間：2019～2022

課題番号：19K20394

研究課題名（和文）大人数の遺伝子発現データ無しで表現型と遺伝子発現の個人差の関連を調べる手法の開発

研究課題名（英文）Development of a method to examine the relationship between phenotype and individual differences in gene expression without the need for gene expression data from a large number of people.

研究代表者

尾崎 遼 (Ozaki, Haruka)

筑波大学・医学医療系・准教授

研究者番号：10743346

交付決定額（研究期間全体）：（直接経費） 3,100,000円

研究成果の概要（和文）：遺伝子発現の個人差と表現型の関連を調べる手法の開発を目指した。その中で、大規模エピゲノムデータに基づき、ゲノム配列の個人差が転写因子結合に与える影響の予測に成功した。また、遺伝子発現変動に生体内環境が与える影響を調べるため、臓器由来の1細胞空間トランスクリプトームデータにおいて各細胞の遺伝子発現変動に周囲の細胞が与える影響を評価した。さらに、遺伝子発現の個人差と表現型の関連について知見を得るため、大規模コホートのがん組織由来バルクRNA-seqデータと少人数のがん組織由来1細胞RNA-seqデータから、発現量の多寡が生存時間に影響を与える細胞間相互作用を抽出し、予後との関連を明らかにした。

研究成果の学術的意義や社会的意義

ヒト疾患関連変異の大半はタンパク質のアミノ酸配列を変えない非コード変異である。これらの非コード変異は主に遺伝子発現量の違いを通じて表現型の違いを生み出すため、さまざまな臓器での発現量の個人差が表現型とどのように関連するかを調べることは、疾患発症リスクや精密医療の観点から重要である。本研究では特に、ゲノムの個人差と転写因子結合の関係や遺伝子発現変動と細胞間相互作用の関係を明らかにすることで、大量の個人ゲノムデータが蓄積しつつある現代社会において、ヒト疾患研究の基盤を与えると期待される。

研究成果の概要（英文）：We aimed to develop a method to investigate the relationship between individual differences in gene expression and phenotype. In this context, we succeeded in predicting the effects of individual differences in genome sequence on transcription factor binding based on large-scale epigenomic data. In addition, to investigate the effect of the in vivo environment on gene expression variation, we evaluated the effect of surrounding cells on gene expression variation in each cell in organ-derived single-cell spatial transcriptome data. Furthermore, to gain insight into the relationship between individual differences in gene expression and phenotype, we extracted cell-cell interactions in which high or low expression levels affect survival time from bulk RNA-seq data derived from a large cohort of cancer tissues and single-cell RNA-seq data derived from a small number of cancer tissues to determine the relationship with prognosis.

研究分野：生命情報科学

キーワード：機械学習 遺伝子発現 深層学習 表現型予測 ゲノム配列 シングルセルRNA-seq

## 1. 研究開始当初の背景

疾患のリスクや形質といった表現型の個人差がゲノム（遺伝型）の違いからどう生み出されているかを明らかにすることは生命科学の究極の目的の一つであると同時に、疾患の予測・予防・治療においても重要な課題である。

ヒトの全身の細胞は、同一のゲノムを持ちつつ、互いに異なる遺伝子発現を示す多様な細胞種に分化し、組織や臓器を形成する。ヒトの疾患との関連が報告されている遺伝的変異の 95%はたんぱく質のアミノ酸配列を変えない「非コード変異」である (Eicher et al., *Nucleic Acids Res.*, 2015)。非コード変異は主に遺伝子発現量の違いを通じて表現型の違いを生み出す。そのため、遺伝的変異と表現型の関連を直接調べるのではなく、特定の臓器・組織・細胞種における遺伝子発現と表現型の関連を調べるトランスクリプトームワイド関連解析 (TWAS) が開発された (Gusev et al., *Nat. Genet.*, 2016)。

TWAS では、表現型の関連をどの組織・細胞種の遺伝子発現と調べるかが結果に大きく影響する。にもかかわらず、TWAS の適用は一部の組織・細胞種に限られてきた。これは、大人数の組織・細胞型ごとの遺伝子発現量データが必要となる一方で、ヒトでは多くの組織・細胞型についてサンプル収集が難しいため、個人のゲノムデータと比べて個人のトランスクリプトームデータは収集が困難である。

実際、疾患サンプルについては手術で採取された検体や患者集団からの死後検体以外は収集が難しく、また、健康なヒトの組織のサンプルは末梢血単球細胞など一部の細胞種に限られる。このため、大人数のサンプル取得が困難な組織・細胞種での遺伝子発現が表現型と関連していたとしても、TWAS を適用できず、表現型との関連を見逃してしまうという問題があった。

遺伝統計学分野と並行して、近年、ゲノム科学の研究者によって、大人数のゲノム・遺伝子発現データを必要とせずに、DNA 配列を入力として遺伝子発現や転写因子結合、エピゲノム修飾といった分子表現型を予測する研究が精力的に取り組まれている。

このアプローチは、(個人間の遺伝型の違いを学習するのではなく) 単一のゲノム配列中の異なる箇所の配列の違いから、DNA 配列と分子表現型の関連を学習する。

申請者も独自に開発したソフトウェアにより、転写因子 ChIP-seq データ (転写因子結合部位を網羅的に測定したデータ) から塩基配列の変異と転写因子結合や転写活性化能の違いが関連することを見出した (Yoshitane et al, *Mol. Cell Biol.*, 2015; Ozaki et al., *Comput. Biol. Chem.*, 2016)。

特に最近では、深層学習も含む機械学習技術が分子表現型の予測に用いられるようになってきた。このように学習されたモデルは、単一のゲノム配列のみを用いているにも関わらず、遺伝子発現への影響が既知の遺伝的変異を導入した新たな DNA 配列を入力にした際には、変異の遺伝子発現への影響を精度よく予測できることが示されている (Kelly and Reshef, *Genome Research*, 2018; Zhou et al., *Nature Genetics*, 2018)。これは、各遺伝子の転写制御領域の配列がある程度共通性を持つことで、特定の組織・細胞種で似た遺伝子発現パターンを示すことから考えて妥当である。

## 2. 研究の目的

本研究では、TWAS の課題であった「大人数の遺伝子発現データ」を必要とせずに、あらゆる組織・細胞型において遺伝子発現と表現型の関連を調べる方法を提案する。提案手法を適用することで、これまで不可能であった組織・細胞型と表現型との組み合わせでの関連解析が可能になり、疾患のリスク予測の正確性向上、疾患の作用機序の発見につながると期待される。本研究では、遺伝統計学分野で研究されてきた TWAS と、機械学習分野で研究されてきたゲノム配列からの遺伝子発現変化予測技術とを組み合わせることで達成される。

## 3. 研究の方法

米国 NIH dbGaP に所収されているヒト個人由来のゲノムデータを所属機関において、関係部局と協議を進め、データ利用に先立ち、倫理審査委員会の承認を得た。また、異なるコホート由来のバルク RNA-seq データを統合して解析する場合に、バッチ効果が問題になるため、補正が必要となる。その中で、1 細胞 RNA-seq 用解析フレームワークの Seurat パイプラインの使用によってバルク RNA-seq データのバッチ効果が軽減されることを見出し、実際の個人由来遺伝子発現データの処理に活用できる知見を得た。

遺伝子発現の個人差と表現型の関連を調べる中で、ゲノムの個人差と転写因子結合の関係を明らかにすることを目指した。ChIP-seq データの二次データベース ChIP-Atlas から 10,000 件以上のヒト転写因子 ChIP-seq データを取得し、独自のサンプルフィルタリングにより、約 3000 件のデータを解析に供した。

遺伝子発現変動に生体内環境が与える影響を調べるため、遺伝子発現の個人差と表現型の関連について知見を得るため、膵臓癌について、3 種類の大規模コホートのがん組織由来バルク RNA-seq データを国際がんゲノムコンソーシアムのデータレポジトリより、少人数のがん組織由

来 1 細胞 RNA-seq データを Gene Expression Omnibus より収集した。

細胞間相互作用が遺伝子発現変動に与える影響の推定のため、1 細胞トランスクリプトーム手法 seq-scope と seqFISH+ の公開データを、データに紐づいた文献の Data Availability セクションの記載に基づき取得した。

#### 4 . 研究成果

遺伝子発現の個人差と表現型の関連を調べる手法の開発を目指した。その中で、(1) ゲノム配列の個人差が転写因子結合に与える影響の予測、(2) 細胞間相互作用に着目した、遺伝子発現と表現型との関連の探索、(3) 細胞間相互作用が遺伝子発現変動に与える影響の推定、について特に研究を進めた。

ゲノム配列の個人差が転写因子結合に与える影響の予測のため、大規模エピゲノムデータに基づき、ゲノム配列の個人差が転写因子結合に与える影響の予測手法を開発した。具体的には、短い塩基配列ごとに転写因子結合特異性スコア (MOCCS スコア) を計算するとともに、1 塩基多型 (SNP) による MOCCS スコアの差分 (MOCCS スコア) を算出した。この MOCCS スコアの絶対値から転写因子結合への予測をおこなった。複数のアッセイデータによる精度検証から、MOCCS スコアによる SNP の転写因子結合への予測は高い精度を示すことが確認された。さらに、複数の種類の疾患のゲノムワイド関連解析から得られた疾患関連 SNP の転写因子結合への影響予測にも成功した。以上の成果は複数の国内学会等で発表し、ポスター賞を受賞した (Tahara, IIBMP2020, 2020-09)。また、プレプリントとして *bioRxiv* に投稿し、現在国際学術雑誌に投稿中である。

さらに、遺伝子発現の個人差と表現型の関連について知見を得るため、大規模コホートのがん組織由来バルク RNA-seq データと少数のがん組織由来 1 細胞 RNA-seq データから、発現量の多寡が生存時間に影響を与える細胞間相互作用を抽出し、予後との関連を調べる手法を開発した。これを膵臓癌由来のデータに適用したところ、予後と関連するリガンド-受容体ペアが複数検出されるとともに、これらのリガンド-受容体ペアがどの細胞型からどの細胞型への細胞間コミュニケーションを担っているのかを推定することができた。以上の成果は、査読付き国際学術雑誌に採択された (Suzuki et al., 2021)。

また、遺伝子発現変動に生体内環境が与える影響を調べるため、臓器由来の 1 細胞空間トランスクリプトームデータにおいて各細胞の遺伝子発現変動に周囲の細胞が与える影響を推定する手法を開発した。この手法では、部分的最小二乗回帰 (partial least squares regression) を援用することで各細胞の遺伝子発現変動に周囲の細胞が与える影響を推定できる。脳組織および肝臓組織由来の 1 細胞空間トランスクリプトームデータに適用したところ、先行知見に合致した細胞間相互作用が抽出されることが確認された。以上の成果は、査読付き国際学術雑誌に採択された (Tsuchiya et al., 2022)。

加えて、上記で得られたデータ処理・データ解析手法・知見を活かし、ゲノム配列・遺伝子発現・表現型の関連を共同研究において貢献することができた (Wakida et al., 2020; Kokaji et al., 2020; Egami et al., 2021; Minoshima et al., 2021; Kokaji et al., 2022)。

ヒト疾患関連変異の大半はタンパク質のアミノ酸配列を変えない非コード変異である。これらの非コード変異は主に遺伝子発現量の違いを通じて表現型の違いを生み出すため、さまざまな臓器での発現量の個人差が表現型とどのように関連するかを調べることは、疾患発症リスクや精密医療の観点から重要である。本研究では特に、ゲノムの個人差と転写因子結合の関係や遺伝子発現変動と細胞間相互作用の関係を明らかにすることで、大量の個人ゲノムデータが蓄積しつつある現代社会において、ヒト疾患研究の基盤を与えると期待される。

## 5. 主な発表論文等

〔雑誌論文〕 計8件（うち査読付論文 7件/うち国際共著 0件/うちオープンアクセス 8件）

1. 著者名 Suzuki Sayaka R., Kuno Akihiro, Ozaki Haruka	4. 巻 28
2. 論文標題 Cell-to-cell interaction analysis of prognostic ligand-receptor pairs in human pancreatic ductal adenocarcinoma	5. 発行年 2021年
3. 雑誌名 Biochemistry and Biophysics Reports	6. 最初と最後の頁 101126 ~ 101126
掲載論文のDOI (デジタルオブジェクト識別子) 10.1016/j.bbrep.2021.101126	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 Minoshima Fumi, Ozaki Haruka, Odaka Haruki, Tateno Hiroaki	4. 巻 24
2. 論文標題 Integrated analysis of glycan and RNA in single cells	5. 発行年 2021年
3. 雑誌名 iScience	6. 最初と最後の頁 102882 ~ 102882
掲載論文のDOI (デジタルオブジェクト識別子) 10.1016/j.isci.2021.102882	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 Wakida, H., Kawata, K., Yamaji, Y., Hattori, E., Tsuchiya, T., Wada, Y., Ozaki, H., & Akimitsu, N.	4. 巻 527
2. 論文標題 Stability of RNA sequences derived from the coronavirus genome in human cells	5. 発行年 2020年
3. 雑誌名 Biochemical and Biophysical Research Communications	6. 最初と最後の頁 993 ~ 999
掲載論文のDOI (デジタルオブジェクト識別子) 10.1016/j.bbrc.2020.05.008	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 Egami, R., Kokaji, T., Hatano, A., Yugi, K., Eto, M., Morita, K., Ohno, S., Fujii, M., Hironaka, K.-I., Uematsu, S., Terakawa, A., Bai Y., Pan, Y., Tsuchiya, T., Ozaki, H., Inoue, H., Uda, S., Kubota H., Suzuki, Y., Matsumoto M., Nakayama, K.-I., Hirayama, A., Soga, T., & Kuroda, S.	4. 巻 24
2. 論文標題 Trans-omic analysis reveals obesity-associated dysregulation of inter-organ metabolic cycles between the liver and skeletal muscle	5. 発行年 2021年
3. 雑誌名 iScience	6. 最初と最後の頁 102217 ~ 102217
掲載論文のDOI (デジタルオブジェクト識別子) 10.1016/j.isci.2021.102217	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 Kokaji, T., Hatano, A., Ito, Y., Yugi, K., Eto, M., Morita, K., Ohno, S., Fujii, M., Hironaka, K.-I., Egami, R., Terakawa, A., Tsuchiya, T., Ozaki, H., Inoue, H., Uda, S., Kubota, H., Suzuki, Y., Ikeda, K., Arita, M., Matsumoto, M., Nakayama, K.-I., Hirayama, A., Soga, T., & Kuroda, S.	4. 巻 13
2. 論文標題 Transomics analysis reveals allosteric and gene regulation axes for altered hepatic glucose-responsive metabolism in obesity	5. 発行年 2020年
3. 雑誌名 Science Signaling	6. 最初と最後の頁 eaaz1236 ~ 1236
掲載論文のDOI (デジタルオブジェクト識別子) 10.1126/scisignal.aaz1236	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 Kokaji T., Eto M., Hatano A., Yugi K., Morita K., Ohno S., Fujii M., Hironaka K., Ito Y., Egami R., Uematsu S., Terakawa A., Pan Y., Maehara H., Li D., Bai Y., Tsuchiya T., Ozaki H., Inoue H., Kubota H., Suzuki Y., Hirayama A., Soga T., Kuroda S.	4. 巻 12
2. 論文標題 In vivo transomic analyses of glucose-responsive metabolism in skeletal muscle reveal core differences between the healthy and obese states	5. 発行年 2022年
3. 雑誌名 Scientific Reports	6. 最初と最後の頁 -
掲載論文のDOI (デジタルオブジェクト識別子) 10.1038/s41598-022-17964-9	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 Tsuchiya Takaho, Hori Hiroki, Ozaki Haruka	4. 巻 38
2. 論文標題 CCPLS reveals cell-type-specific spatial dependence of transcriptomes in single cells	5. 発行年 2022年
3. 雑誌名 Bioinformatics	6. 最初と最後の頁 4868 ~ 4877
掲載論文のDOI (デジタルオブジェクト識別子) 10.1093/bioinformatics/btac599	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 Tahara Saeko, Tsuchiya Takaho, Matsumoto Hirotaka, Ozaki Haruka	4. 巻 -
2. 論文標題 MOCCS profile analysis clarifies the cell type dependency of transcription factor-binding sequences and cis-regulatory SNPs in humans	5. 発行年 2022年
3. 雑誌名 bioRxiv	6. 最初と最後の頁 -
掲載論文のDOI (デジタルオブジェクト識別子) 10.1101/2022.04.08.487641	査読の有無 無
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 -

〔学会発表〕 計9件（うち招待講演 1件 / うち国際学会 0件）

1. 発表者名 尾崎遼
2. 発表標題 NGSデータ解析における生命現象のモデル化と自動化
3. 学会等名 バイオDXの最前線 - JST戦略的創造研究推進事業 CREST「データ駆動・AI駆動を中心としたデジタルトランスフォーメーションによる生命科学の革新[バイオDX]」キックオフシンポジウム（招待講演）
4. 発表年 2021年

1. 発表者名 Saeko Tahara, Takaho Tsuchiya, Haruka Ozaki
2. 発表標題 ChIP-seqデータベースの大規模解析で解明する細胞型ごとに多彩な転写因子認識配列
3. 学会等名 IIBMP2020
4. 発表年 2020年

1. 発表者名 田原沙絵子, 土屋貴穂, 尾崎遼
2. 発表標題 ChIP-Atlasの二次解析による多彩な転写因子認識配列の解明および更なるデータベース化
3. 学会等名 トーゴの日シンポジウム2020
4. 発表年 2020年

1. 発表者名 Saeko Tahara, Takaho Tsuchiya and Haruka Ozaki
2. 発表標題 k-mer profile analysis clarifies cell-type specific TF binding landscape in human
3. 学会等名 2021年日本バイオインフォマティクス学会年会・第10回生命医薬情報学連合大会（IIBMP2021）
4. 発表年 2021年

1. 発表者名 Takaho Tsuchiya, Haruka Ozaki
2. 発表標題 Cell-Cell PLS regression analysis reveals cell-type specific spatial dependency of transcriptome in single cells
3. 学会等名 2021年日本バイオインフォマティクス学会年会・第10回生命医薬情報学連合大会 (IIBMP2021)
4. 発表年 2021年

1. 発表者名 土屋貴穂, 尾崎遼
2. 発表標題 CCPLS reveals cell-type specific spatial dependence of transcriptome in single cells
3. 学会等名 バイオDXの最前線 - JST戦略的創造研究推進事業 CREST「データ駆動・AI駆動を中心としたデジタルトランスフォーメーションによる生命科学の革新[バイオDX]」キックオフシンポジウム
4. 発表年 2021年

1. 発表者名 Saeko Tahara, Takaho Tsuchiya, Hirotaka Matsumoto, Haruka Ozaki
2. 発表標題 MOCCSプロファイルで明らかにする、転写因子認識配列の細胞型特異性とヒト一塩基多型が転写因子結合へ与える影響 (MOCCS profile analysis clarifies the cell type dependency of transcription factor-binding sequences and cis-regulatory SNPs in humans)
3. 学会等名 2022年日本バイオインフォマティクス学会年会・第11回生命医薬情報学連合大会 (IIBMP2022)
4. 発表年 2022年

1. 発表者名 Hirofumi Kariyayama, Takeshi Kawashima, Haruka Ozaki, Hiroshi Wada
2. 発表標題 円口類ゲノムを利用したドメインシャッフリングによって獲得された新規遺伝子の探索
3. 学会等名 2022年日本バイオインフォマティクス学会年会・第11回生命医薬情報学連合大会 (IIBMP2022)
4. 発表年 2022年

1. 発表者名 田原沙絵子, 土屋貴穂, 松本拡高, 尾崎遼
2. 発表標題 Predicting SNPs' effects on transcriptional factor binding and cell-type specificity using a crowd of ChIP-seq data
3. 学会等名 日本人類遺伝学会第67回大会
4. 発表年 2022年

〔図書〕 計0件

〔産業財産権〕

〔その他〕

ポスター受賞. Saeko Tahara, Takaho Tsuchiya, Haruka Ozaki. ChIP-seqデータベースの大規模解析で解明する細胞型ごとに多彩な転写因子認識配列. IIBMP2020. 2020-09
---

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

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------