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

平成　３０　年　　５　月　２３　日現在

機関番号：　１２６０１
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ１８４６５

研究課題名（和文）エピゲノム比較解析パイプライン高度化のための正規化・統合解析手法の構築

研究課題名（英文）Integration and normalization method for improving the comparative epigenome pipeline

研究代表者

中戸　隆一郎（Nakato, Ryuichiro）

東京大学・分子細胞生物学研究所・助教

研究者番号：６０５８３０４４

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：多数のChIP-seqサンプルを比較する大規模ChIP-seq解析では、細胞特異的なタンパク結合や、DNA結合におけるタンパク間の依存関係を捉えることができる。しかし微量細胞や生体組織など困難な条件下で調整されたサンプルはデータに強い偏りが発生する場合があり、従来法ではデータ補正も困難なため、現状では比較解析に含めることができない。本申請ではこれらの課題を克服すべく、効率的に同時解析可能な解析プログラムDROMPA3、新たなChIP-seq品質評価手法SSPおよび、Hi-Cデータ解析パイプラインを作成した。これらの新規手法はより頑健かつ効率的な大規模エピゲノム解析を可能にする。

研究成果の概要（英文）：Recent advances in sequencing analyses enable us to compare hundreds of ChIP-seq samples simultaneously; such large-scale analysis has potential to reveal the high-dimensional interrelationship level for regulatory elements and annotate novel functional genomic regions de novo. However, there are various factors that can affect the data quality of the sample preparation step, especially for tissue samples and low-input analyses. It is still difficult to eliminate or normalize biases in each sample. To overcome this problem, we have developed three programs: (i) DROMPA3, cost-effective ChIP-seq pipeline; (ii) SSP, a novel quality assessment tool of ChIP-seq data; and (iii) Hi-C analysis pipeline. These programs can provide us with more robust and effective large-scale epigenome analysis.

研究分野：　バイオインフォマティクス

キーワード：　ChIP-seq法　Hi-C法　大規模解析　エピゲノム　品質評価　ゲノム立体構造

１．研究開始当初の背景

ChIP-seq 法はゲノム上のタンパク質結合部位およびヒストン修飾部位を網羅的に収集する手法である。複数の ChIP-seq サンプルを解析し得られたエピゲノムプロファイルを異なる細胞種間で比較することで、数々の病態を規定する細胞特異的な遺伝子発現制御機構や、ゲノム上に埋め込まれた未知の機能性制御領域を捉えることが可能となる。代表的な例として、117 種もの細胞種について種々のヒストン修飾データを網羅的に収集し、ゲノムの各領域を 15 種類のクロマチン状態(Chromatin state) に分類・比較した大規模解析が報告されている[ROADMAP 2015]。しかしながら、そのような大規模解析には、以下のような困難な問題が存在し、実現可能な場面は限られているのが現状である。

(1) 大規模解析では、データが十分高品質でありかつ複数の複製（replicate）が用意できることを暗黙の前提としているが、ChIP-seq データの品質はサンプル調製上の種々の要因（抗体のロット、タンパク質固定、DNA 断片化、PCR 増幅など）の影響を強く受けるため、ばらつき（ノイズ）が起きやすく、サンプルの再生産コストが高い。また、解析が大規模化し実験が長期化した場合、実験条件を同一に維持し続けることは概して困難である。

(2) 生体細胞（患者細胞など）を用いた実験では個人差によるばらつきを除外するために複製数を増やす必要があるが、１検体につき 1 サンプルで済む RNA-seq や DNA メチル化と異なり、エピゲノム解析は１検体につき複数の（通常 5 種以上）ChIP-seq サンプルを取得する必要があり、うち１つでもデータが低品質として欠けてしまうと、そのデータセット全体が解析に利用できなくなる。

(3) 微量細胞など困難な条件下で調整されたサンプルはデータに強い偏り（ノイズ）がしばしば発生する一方、何度もデータを再生産することはコスト面で難しい。

２．研究の目的

本研究では上記の課題を克服し、従来法では比較解析に含めることができなかった最高品質でないサンプルも可能な限り許容しつつ、ノイズを含めた大量のプロファイルから信頼性の高い結果のみを抽出する頑健な頑健かつ効率的なエピゲノム解析パイプラインを構築することを目標とする。これにより、大規模解析に必要なサンプル生成コスト、適用範囲を大幅に低減・拡大させられるのみならず、得られる結果の信頼性も向上させる。開発したプログラム、得られた知見は積極的に公開し、国内外の ChIP-seq 解析研究に貢献する。

３．研究の方法

本研究は以下の課題を設定して進めた。

(1) 偏りを持つサンプルからのノイズ除去・正規化手法の開発

(2) 既存の品質評価指標（リード数、サンプル冗長度、GC 含量など）で検出できない低品質サンプルを検出するための新たな評価指標の考案

(3) 複数の ChIP-seq サンプルを入力としたエピゲノムプロファイル可視化・アノテーションプログラムの開発

開発したプログラムを国際ヒトエピゲノムコンソーシアム（IHEC, ihec-epigenomes.org）プロジェクトにおいて生産されたヒト血管内皮細胞データに適用し、新規知見を獲得する。

４．研究成果

(1) Spike-in 解析を用いたノイズ除去法

Spike-in 解析とは、ChIP-seq 比較解析の際に異なる生物種由来のゲノム DNA を全サンプルに等量入れ、internal control として正規化に用いることで、通常の解析では検出不可能なピーク強度の絶対的変化を測定する手法である。我々のコヒーシン病(CdLS)患者細胞を用いた ChIP-seq 実験では、得られるデータにしばしば強い GC 含量の偏りが起き、qPCR では減少しているピークが ChIP-seq では逆に増加して見える場合があった。Spike-in 解析による検証の結果、比較するサンプル間で目的タンパク質の DNA 結合量が大きく異なる場合、免疫沈降によって得られた DNA 量に含まれるバックグラウンド（非特異的）リードの割合が大きく異なっており、バックグラウンド量がサンプル間で一定であると暗黙に仮定する従来の正規化手法で過度に増幅されたバックグラウンドがノイズとして現れることがわかった。これらの偏りは PCR 増幅を行う前の初期 DNA 量が少ない場合及び、DNA 断片化処理に DNase を用いた場合などに起きやすいことも併せて判明した。

(2) 多サンプル比較のための正規化手法の構築

従来のサンプル間正規化手法は、現実にしばしば起きうるサンプル間の S/N 比のばらつきを考慮していない。そこで本研究では Quantile 正規化によるピーク強度補正法を開発した。本手法では比較するサンプル間で共通するピーク領域でのピーク強度分布を基準に正規化し（図１上）、同一抗体を用いたサンプルの比較の際に S/N 比のばらつきに対して頑健な定量的比較を可能とする。IHEC プロジェクトで得られたヒト血管内皮細胞のヒストン修飾データに対しこの手法を用いて組織間サンプルクラスタリングを行った結果、良好なクラスタリング結果を得ることができ(図１下)、各内皮特異的な発現変動遺伝子及びプロモーター・エンハンサー領域の同定に成功した。この成果について現在論文投稿準備中である。

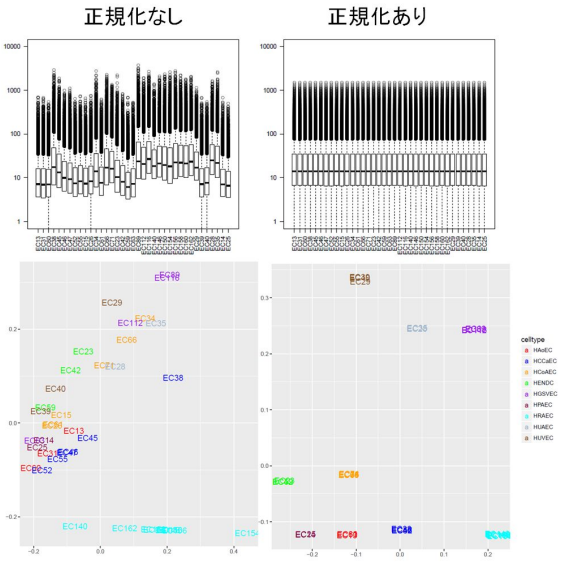

図１：Quantile 正規化。（上）サンプルの S/N 比の箱ひげ図。（下）クラスタリングの例。正規化によって同じ色で示された同一細胞種の複製がクラスタ化されている。

(3) 新規の品質評価手法の考案

IHEC プロジェクトの大規模解析を進めていくなかで、従来の ChIP サンプル品質評価手法に問題があることが明らかになった。すなわち、H3K9me3 ヒストン修飾など広範囲に修飾が分布するサンプルの S/N 比を評価できない、値がサンプルのマップリード数に依存する、得られたピークの信頼性（ノイズをどの程度含むか）を評価できない点である。そのため、解析に用いて良いサンプルと棄却すべきサンプルの客観的な区別ができない状態にあった。これらの問題点を克服すべく、品質評価のための新規手法"SSP"を開発した。本手法は、ゲノムの順鎖・逆鎖それぞれにマップされたリードの相関の強さを Jaccard index を利用して計測し、strand-specific profile を描画することで、リード数・細胞種・生物種によらず統一的に S/N 比を評価することができる。また、バックグラウンド領域におけるマップリードのばらつきを定量化した Background uniformity という指標を新たに考案し、従来のサンプル冗長度や GC 含量などでは検出できなかった低品質サンプルを検出することを可能とした。この成果は Bioinformatics 誌に発表されている[Nakato 2018]。本論文はプレプリントサーバ bioRxiv にも併せて投稿しており、オープンアクセスで誰でも閲覧可能である(doi: 10.1101/165050)。

(4) 立体構造解析パイプラインの構築

エピゲノム解析で得られる情報の信頼度と価値をさらに高めるため、Hi-C 法、ChIA-PET 法のための立体構造データ解析のパイプラインを構築した。本パイプラインにより、Hi-C データから得られる染色体の大きな二分構造（コンパートメント）、立体相互作用のまとまり（TAD）、エンハンサー・プロモーター間相互作用などを獲得できる。これらを ChIP-seq 解析で得られるエピゲノム情報と統合し、遠位のエンハンサーと遺伝子の紐づけ、ヒストン修飾境界の高精度な同定、さまざまなエピゲノム状態と立体総合作用頻度との関連など、いくつかの新規な知見が得られている。この成果について現在論文投稿準備中である。

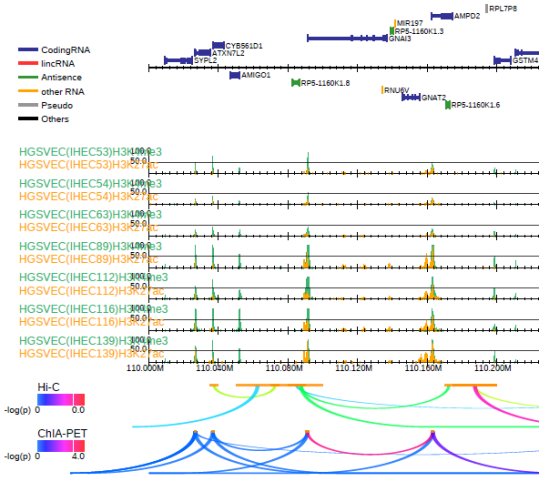

図２：DROMPA による可視化例。上段に遺伝子、中段に ChIP-seq プロファイル（緑：プロモーター、橙：エンハンサー）、下段に Hi-C 及び ChIA-PET の立体相互作用を表示している。

(5) 複数サンプルを統合的に解析するエピゲノムプロファイル可視化・アノテーションプログラム DROMPA

得られたエピゲノムデータ、立体構造データなどを効率的に解析するプログラム DROMPA3 を開発した(図２)。本手法は上述した Spike-in 解析、ノイズ除去、正規化、比較解析などに対応しており、さまざまな実験データ・用途に対して効率的に適用可能である。さらに、多サンプルを用いた大規模解析を進めていく中で培った知見をまとめた総説を Briefing in Bioinformatics 誌に発表した[Nakato 2017]。本総説は最新の手法や知見を体系立てて紹介する ChIP-seq 解析ガイドラインとして国内外で広く活用されている。

(6) ブログ開設

開発した研究成果を日本社会に還元していくにあたり、英語で書かれた論文やマニュアルは浸透度が低く、似た質問を繰り返し受けるようになったことから、日本語で ChIP-seq 解析を解説した Web ページの必要性を痛感した。そこで、申請者の開発した DROMPA, SSP を含めた ChIP-seq 解析のノウハウを解説するブログを開設した(rnakato.hatenablog.jp/)。ChIP-seq 解析に不慣れな初心者や学生にわかりやすくする

ため平易な日本語で記述することに重きを置いており、これまでのところ好評を得ていると共に、開発した DROMPA, SSP **の普及にも役立っている。本ブログは本課題終了後も引き続き更新を続ける予定である。**

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計５件）

1. Nakato R, Shirahige K, Sensitive and robust assessment of ChIP-seq read distribution using a strand-shift profile, Bioinformatics, Epub ahead of print, **査読有**, doi: 10.1093/bioinformatics/bty137, (2018)
2. Nakato R, Shirahige K, Statistical Analysis and Quality Assessment of ChIP-seq Data with DROMPA. Methods in Molecular Biology, 1672, 631-643, **査読無**, doi: 10.1007/978-1-4939-7306-4_41, (2018)
3. Nakato R, Shirahige K, Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation, Briefings in Bioinformatics, vol. 18, issue 2, 279-290, **査読有**, doi: 10.1093/bib/bbw023, (2017)
4. Sutani T, Sakata T, Nakato R., Masuda K, Ishibashi M, Yamashita D, Suzuki Y, Hirano T, Bando M, Shirahige K., Condensin targets and reduces unwound DNA structures associated with transcription in mitotic chromosome condensation, Nature Communications, vol. 6, issue 7815, 1-13, **査読有**, doi: 10.1038/ncomms8815, (2015)
5. Minamino M., Ishibashi M., Nakato R., Akiyama K., Tanaka H., Kato Y., Negishi L., Hirota T., Sutani T., Bando M., Shirahige K., Esco1 Acetylates Cohesin via a Mechanism Different from That of Esco2, Current Biology, vol. 25, issue 13, 1694-1706, **査読有**, doi: 10.1016/j.cub.2015.05.017, (2015)

〔学会発表〕（計　５件）

1. Nakato R., Sakata T, Shinkai S, Shirahige K., Characterization of chromatin folding mechanisms regulated by cohesin, cohesin loader and CTCF, Keystone Symposia (2018).
2. Nakato R. and Shirahige K., Method of identifying biased data for large-scale ChIP-seq analysis, The EMBO Meeting "From Functional Genomics to Systems Biology" (2016).
3. Nakato R., Wada Y., Katou Y., Nakaki R., Tsutsumi S., Mituyama T., Kimura H., Aburatani H., and Shirahige K., Integrative epigenome profiling reveals tissue-specific gene regulations of human vascular endothelial cells, Cold Spring Harbor Laboratory (CSHL) 2016 meetings "Systems Biology" (2016).
4. **中戸隆一郎**, **バイオインフォマティクスで切り開くエピゲノム解析**, Heart rhythm interdisciplinary Association Forum 2017 (2017).
5. **中戸隆一郎**, **ゲノムのダイナミズムを解き明かす大規模** ChIP-seq **解析**, **理研シンポジウム「細胞システムの動態と論理** IX」(2017).

〔図書〕（計２件）

1. **中戸隆一郎**, **エピゲノム解析最前線** —ChIP-seq **解析法の現在と未来**, **医学書院「生体の科学」**68 **巻** 3 **号**, pp. 204-208 (2017).
2. **中戸隆一郎**, **和田洋一郎**, **白髭克彦**, **エピゲノムのダイナミズムを解き明かす大規模比較解析　—血管内皮細胞を例に**, **羊土社「実験医学」**2016 **年** 6 **月増刊号**, P1562-1568 (2016).

〔産業財産権〕
○出願状況（計０件）
○取得状況（計０件）

〔その他〕
DROMPA3:
https://github.com/rnakato/DROMPA3

SSP: https://github.com/rnakato/SSP

Blog: http://rnakato.hatenablog.jp/

６．研究組織
(1)研究代表者
中戸　隆一郎（NAKATO RYUICHIRO）
東京大学・分子細胞生物学研究所・助教
研究者番号：60583044