

科学研究費助成事業 研究成果報告書

平成 27 年 5 月 25 日現在

機関番号：11301

研究種目：若手研究(B)

研究期間：2013～2014

課題番号：25730011

研究課題名(和文) エピゲノム及び遺伝子による細胞内制御ネットワークモデリングと細胞分化機構の解析

研究課題名(英文) Modeling of genes and epigenome intracellular regulatory network and analysis of cellular differentiation mechanisms

研究代表者

小島 要 (Kojima, Kaname)

東北大学・東北メディカル・メガバンク機構・講師

研究者番号：10646988

交付決定額(研究期間全体)：(直接経費) 1,700,000円

研究成果の概要(和文)：細胞内システムの解析において、遺伝子発現量のみを考慮した制御関係ネットワークが考えられてきたが、細胞分化の過程において遺伝子だけでなくDNAメチル化やヒストン修飾等のエピゲノム状態の変化により、制御関係が変化し、異なる機能を持つ細胞へ分化すると考えられている。本研究では、遺伝子とエピゲノムによる制御ネットワークの解析に用いられる、遺伝子発現量の高精度な推定手法の開発、並びにエピゲノム状態推定の高精度化のための基盤手法の開発を行う。さらに、細胞内制御ネットワークの解析を目的として、マウスの細胞分化過程における遺伝子発現量データならびにヒストン修飾データからの転写因子の結合予測手法の開発を行う。

研究成果の概要(英文)：Regulatory networks comprised of only genes have been considered for the usual analysis of intracellular system. In the cell differentiation process, the alteration of epigenetic states such as DNA methylation and histone modification induces the change of regulatory structure and the change enables cells to be differentiated to other cells with difference functions. In this study, we devise new methods for estimating the abundance of gene expression, and methods for the foundation of highly accurate estimation of epigenetic states. Targeting the analysis of intracellular regulatory network, we devise the approach for estimating cell-specific transcription factor binding from gene expression data and histone modification data in the cell differentiation process of mouse.

研究分野：バイオインフォマティクス

キーワード：細胞内制御ネットワーク エピジェネティクス 次世代シーケンサー

1. 研究開始当初の背景

マイクロアレイ技術の登場により、細胞内における数千から数万の遺伝子の発現量の計測が可能となり、いくつかの遺伝子をノックダウンした後に、遺伝子発現量を計測したノックダウン遺伝子発現データや、薬剤応答等に対する遺伝子発現量の時系列変化を捉えるための時系列遺伝子発現データが多く計測されている。これらの遺伝子発現データから、細胞内における生体分子による制御システムの解明を目的として、遺伝子間の制御関係ネットワーク(以下、遺伝子ネットワーク)の推定手法が研究されている(Studying and modelling dynamic biological processes using time-series gene expression data, Bar-Joseph et al., 2012)。

近年、次世代 DNA シークエンシング技術の発展により、ChIP-Seq によるヒストン修飾状態の計測、Bisulfite シークエンシングによる DNA メチル化状態の計測など、全ゲノム的なエピゲノムデータの計測が可能となってきている。ヒストン修飾によるクロマチン構造の変化や DNA メチル化等のエピゲノムデータは、遺伝子の発現の制御、中でも抑制機構に関わっていることが明らかになっており、発現する遺伝子をエピゲノムにより選択的に制御することで皮膚や臓器等の異なる細胞種における機能・形質に違いを与えたと考えられている。このため、細胞内システム、特に細胞分化における機構の解明を目的として、国際ヒトエピゲノムコンソーシアム(<http://ihc-epigenomes.net/>)では、1000 種類のヒトのエピゲノムの解読が目標として掲げられており、ENCODE プロジェクト(<http://encodeproject.org/ENCODE/>)では、ヒトやマウスの個々の細胞種におけるクロマチン構造の状態及び DNA メチル化におけるデータが順次、公開されている。しかしながら、遺伝子のみによる制御ネットワーク推定については、数多くの手法が研究されているものの、エピゲノムデータを網羅的に用いた、制御ネットワーク解析研究は、ほとんど報告されていなかった。

2. 研究の目的

細胞分化過程における遺伝子ならびに DNA メチル化やヒストン修飾などのエピゲノムによる制御機構の解明には、遺伝子発現量とエピゲノム状態データからの制御ネットワーク解析技術が求められる。そこで、上記制御ネットワーク解析のための遺伝子発現を高精度に定量する RNA-Seq 解析技術の開発、並びに次世代シークエン্সデータからの DNA メチル化とヒストン修飾状態の高精度推定のための基盤手法の開発を行う。さらに遺伝子発現量とエピゲノム状態から分化過程の各細胞における転写因子の結合状態を推定することで、全ゲノム的なヒストン修飾によるクロマチン構造の状態、DNA メチル化状態のデータを用いた、エピゲノムと遺伝子による、

高精度細胞内制御ネットワーク解析基盤を提案する。

3. 研究の方法

本研究の方法について下記に 4 項目分けてそれぞれ述べる。

(1) DNA メチル化検出に向けた基盤手法開発

DNA メチル化の検出には Bisulfite シークエン্স法が用いられる。Bisulfite 反応では、メチル化されていないシトシン(C)がウラシル(U)に変化し、シークエン্স結果では、チミン(T)となることから、シトシンからチミンへの変化を検出からメチル化状態が推定される。しかしながら、通常より考慮される塩基の種類が減少するため、より正確なシークエン্স配列のマッピング、変異コールを行う必要があり、リファレンスゲノムにマッピングされたシークエン্সデータからより正確に変異コールを行う手法の開発を行う。

(2) ヒストン修飾状態検出のための ChIP-Seq 精度向上に向けた難読化領域の解析

単純反復配列、コピー数変異、HLA 領域などの難読化領域におけるゲノム情報がヒストン修飾、転写因子の結合情報取得のための ChIP-Seq 解析でのピークコールの高精度化において求められる。そこで、単純反復配列、コピー数変異、HLA 領域の解析手法の開発を行う。

(3) 遺伝子発現定量のための手法開発

遺伝子発現データからの遺伝子制御ネットワークの推定には、遺伝子発現量の計測の精度が重要となる。近年 RNA-Seq 技術により直接 mRNA のシークエンシングを行い、リファレンス配列にマッピングすることで、従来技術よりも精度の高い遺伝子発現量の定量が可能となっている。

(4) エピゲノムデータからの細胞別転写因子結合予測手法の開発

細胞分化過程における遺伝子制御ネットワーク推定を目的とした、エピゲノムデータ及び遺伝子発現量データからの細胞種特有の転写因子の結合推定方法の開発を行う。

4. 研究成果

リファレンスゲノムにマッピングされたシークエン্সデータからより正確に変異コールを行うため、各シークエン্সリードだけでなくリードペアを通じて複数のヘテロ変異をまたぐリード情報を考慮して局所的に変異のフェージングを行いながら、フェージング情報を利用しつつ変異コールを行う統計的手法を開発し、学術論文として発表を行った(雑誌論文、)。また、遺伝子発現データからの遺伝子制御ネットワークの推定には、遺伝子発現量の計測の精度が重要となる。近年 RNA-Seq 技術により直接 mRNA のシークエンシングを行い、リファレンス配列に

マッピングすることで、従来技術よりも精度の高い遺伝子発現量の定量が可能となっている。しかしながら、シーケンシングの際のアライメントにおいてギャップの考慮などが不十分である問題があった。このため、ギャップ付きアライメントを考慮し、各スプライシングバリエーションの単位で正確に遺伝子発現量の定量するための手法を開発し、学術論文として発表した(雑誌論文)。さらにパラメータ更新の枝刈りにより精度を維持したままの高速化を実現している(雑誌論文)。

単純反復配列について、フェージングされた周辺 SNP からの遺伝子系図情報を用いた次世代シーケンシング(NGS)データによる高精度な単純反復配列構造推定手法を提案した(学会発表)。コピー数変異について、既存手法では、似通ったゲノム領域においては、コピー数が合わせて求められる問題があったが、集団の NGS データから LDA を元にしたモデルにより異なるゲノム領域のコピー数として分離・推定する方法を提案した(雑誌論文)。HLA 領域については、昨年度提案の RNA-Seq からの発現量推定法を拡張し、既知 HLA 領域を鋳型とし、NGS データからのリードをマップすることで既存手法より高精度に HLA 型を推定する方法を提案した(雑誌論文)。

マウスにおける ES 細胞から心臓細胞への分化過程における RNA-Seq データ、ヒストン修飾及び転写因子の ChIP-Seq データからのヒストン修飾情報等の特徴量と転写因子結合の正例・負例のデータについて一層目を Gaussian Restricted Boltzmann Machine とし、二層目と三層目でオートエンコーダーを形成する三層の Deep Neural Network で学習し、既存手法より高精度に細胞種特有の転写因子結合情報を推定する手法を提案した。

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7件)

Naoki Nariyai, Kaname Kojima, Sakae Saito, Takahiro Mimori, Yukuto Sato, Yosuke Kawai, Yumi Yamaguchi-Kabata, Jun Yasuda, and Masao Nagasaki, HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data, BMC Genomics, 16(Suppl 2):1-6, 2015. (査読有り)
Takahiro Mimori, Naoki Nariyai, Kaname Kojima, Yukuto Sato, Yosuke Kawai, Yumi Yamaguchi-Kabata, and Masao Nagasaki, Estimating copy numbers of alleles from population-scale high-throughput sequencing data, BMC Bioinformatics, 16(Suppl 1):1-8, 2015. (査読有り)

Naoki Nariyai, Kaname Kojima, Takahiro Mimori, Yukuto Sato, Yosuke Kawai, Yumi Yamaguchi-Kabata, and Masao Nagasaki, TIGAR2: sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads, BMC Genomics, 15(Suppl 9):1-9, 2014. (査読有り)

Tomohiko Ohtuski, Naoki Nariyai, Kaname Kojima, Takahiro Mimori, Yukuto Sato, Yosuke Kawai, Yumi Yamaguchi-Kabata, and Masao Nagasaki, SVEM: a structural variant estimation method using multi-mapped reads on breakpoints, Lecture Notes in Computer Science, 8542:208-219, 2014. (査読有り)

Kaname Kojima, Naoki Nariyai, Takahiro Mimori, Yumi Yamaguchi-Kabata, Yukuto Sato, Yosuke Kawai, and Masao Nagasaki, HapMonster: a statistically unified approach for variant calling and haplotyping based on phase-informative reads, Lecture Notes in Computer Science, 8542:107-118, 2014. (査読有り)

Naoki Nariyai, Osamu Hirose, Kaname Kojima, and Masao Nagasaki, TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference, Bioinformatics, 29(18), 2292-2299, 2013. (査読有り)

Kaname Kojima, Naoki Nariyai, Takahiro Mimori, Mamoru Takahashi, Yumi Yamaguchi-Kabata, Yukuto Sato, and Masao Nagasaki, A statistical variant calling approach from pedigree information and local haplotyping with phase informative reads, Bioinformatics, 29(22):2835-2843, 2013. (査読有り)

[学会発表](計2件)

Kaname Kojima, Yosuke Kawai, Naoki Nariyai, Takahiro Mimori, Takanori Hasegawa, and Masao Nagasaki, Short tandem repeat number estimation from paired-end sequence reads by considering unobserved genealogy of multiple individuals, 11th International Symposium on Bioinformatics Research and Applications, アメリカ合衆国バージニア州ノーフォーク市オールドドミニオン大学、2015年6月9日(査読有り、発表確定)

Kaname Kojima, Statistical methods for analyzing next generation sequencing data, International Statistical

Symposium CSA-KSS-JSS Special Invited
Sessions(招待講演), 台湾、新竹市、交
通大学、2014年12月6日

〔図書〕(計0件)

〔産業財産権〕
出願状況(計0件)

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

取得状況(計0件)

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

6. 研究組織

(1) 研究代表者

小島 要 (KOJIMA, Kaname)
東北大学・東北メディカル・メガバンク機
構・講師
研究者番号：10646988

(2) 研究分担者

()

研究者番号：

(3) 連携研究者

()

研究者番号：