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　元 年　 ６ 月　 ７ 日現在

機関番号：１１３０１
研究種目：若手研究(B)
研究期間：2017～2018
課題番号：１７Ｋ１７５９０

研究課題名（和文）ゲノム情報からのマイクロサテライト統合解析基盤構築による網羅的な疾患関連因子同定

研究課題名（英文）Exhaustive discovery of disease related variants by the construction of integrated analysis methods of microsatellites from genome data

研究代表者
小島　要（Kaname, Kojima）

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）　　1,600,000円

研究成果の概要（和文）：これまでSNPアレイデータからの単塩基変異を主とした関連解析が多くなされている一方マイクロサテライトのリピート数多型に関する研究は限られたものとなっているが、これは既存のマイクロサテライトにおける解析手法の精度が原因と考えられる。本研究では、シークエンスデータからのマイクロサテライトにおけるリピート数推定法と遺伝子系図情報を考慮したリピート数のインピュテーション手法を開発した。さらに開発手法の日本人ゲノムデータへの適用を目的とした日本人全ゲノムシークエンスデータからのマイクロサテライト多型情報の整備と関連解析において重要となる集団構造解析などの知見を考慮した解析パイプラインの構築を行った。

研究成果の学術的意義や社会的意義
マイクロサテライトにおけるリピート数の違いが疾患罹患率に関連することが知られており、ハンチンチン遺伝子領域におけるCAGリピート数増加とハンチントン病の関連などが報告されている。しかしながら、リピート数と疾患などの表現型との関連の多くが未発見であると考えられており、未発見の疾患関連マイクロサテライトの同定にはゲノムデータからのより高精度なリピート数推定手法の開発が重要となる。本研究では既存手法と比べ高精度なゲノムデータからのリピート数推定法を開発しており、開発手法のゲノムデータへの適用と新規疾患関連マイクロサテライトの同定を進めることで疾患発症のメカニズム解明に繋がることが期待される。

研究成果の概要（英文）：There exist a number of association studies for single nucleotide variants on SNP array data. On the other hand, there exist not so many studies for analyzing the association of repeat number polymorphisms on microsatellites due to the limited accuracy of existing methods for analyzing genome data on microsatellites. In this study, we developed a method estimating repeat numbers of microsatellites from sequencing data and an imputation method for repeat numbers of microsatellites in which the genealogy information is explicitly considered. For the purpose of applying our developed methods to Japanese genome data, we also prepared the information of microsatellite polymorphisms in Japanese from whole genome sequencing data, and constructed a pipeline for the association studies that considers the knowledge essential for association studies such as the knowledge for population structure.

研究分野：バイオインフォマティクス

キーワード：マイクロサテライト　遺伝子型インピューテーション　ハイスループットシークエンサー

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

ゲノム配列上において数塩基からなる配列のリピートにより構成される領域はマイクロサテライトと呼ばれ、リピート数の違いが疾患の罹患率に関連することが知られている。例えば、ハンチンチン遺伝子領域における CAG からなる配列のリピート数の増加とハンチントン病の関連やアンドロゲン受容体遺伝子領域における CAG からなる配列のリピート数の増加と球脊髄性筋萎縮症の関連がこれまで報告されている。しかしながら、ゲノム上に点在するマイクロサテライトのリピート数と疾患をはじめとした表現型の関連の多くが未発見であると考えられており、こうしたマイクロサテライトのリピート数と表現型との関連の同定にはゲノムデータからのリピート数推定手法の高精度化と解析検体数の大規模化が必要である。

現在のゲノムデータ計測技術ではハイスループットシークエンサーと SNP アレイの二種類の計測法が主に用いられている。ハイスループットシークエンサーでは数百塩基からなる DNA 配列を文字列化したシークエンスデータが大量に出力される。リシークエンスと呼ばれる解析方法では、ヒトなど各生物種のゲノム配列を代表する参照ゲノムに対してシークエンスデータのアラインメントを行い、アラインメントされたデータに対して単塩基変異をはじめとした変異検出などの網羅的な解析が行われる。しかしながら、依然として一検体当りのコストは高く、ゲノムワイド関連解析などによる症例対照研究を行う際に各症例ごとに数千から数万といった大規模な検体数のゲノムデータを取得することは困難である。このため、シークエンシングと比較してコストが約 10 分の 1 である SNP アレイによりゲノムデータが取得される場合が多い。しかしながら、SNP アレイでは取得できる遺伝子型情報は予め設計された数十万から数百万の単塩基変異や少数の挿入・欠損変異に限られることから、関連解析により同定される疾患関連座位の解像度が低い問題があった。

近年、全ゲノムシークエンスデータから取得された数千検体の変異情報からのハプロタイプにより構成される参照パネルをもとに、SNP アレイで設計された変異の遺伝子型情報から SNP アレイにない変異の遺伝子型情報を復元する遺伝子型インピュテーション法が提案されている。遺伝子型インピュテーション法では組み換えを考慮に入れた形で参照パネル内のハプロタイプを選択することで遺伝子型情報の復元が行われ、この復元された遺伝子型情報を用いることでより高い解像度の関連解析が可能となっている。マクロサテライトについても同様に大規模検体への網羅的な関連解析を実現するためには、SNP アレイにおける遺伝子型情報からのリピート数復元を行うインピュテーション法の利用が望まれる。

２．研究の目的

マイクロサテライト周辺において単位配列のリピートによる配列の特異性の低下と参照ゲノムとのリピート数の違いに起因した挿入・欠損配列により、リシークエンス解析におけるシークエンスデータのアラインメント品質が低下することが分かっている。このため、リシークエンス解析されたシークエンスデータに対してマイクロサテライトの単位配列を事前情報として用いるリアライメント手法を開発しているが（Kojima et al., STR-realigner: a realignment method for short tandem repeat regions, BMC Genomics, 2016）、この知見をもとにマイクロサテライトにおけるアラインメントの誤りを考慮し、補正することでより高精度なマイクロサテライトリピート数の推定を可能とする統計手法を開発する。

マイクロサテライトリピート数のインピュテーションについては、これまで Beagle version 4.1 などの単塩基変異を主に対象としたインピュテーション法をそのまま用いた解析手法が提案されているが、高精度にインピュテーションを行うためにはマイクロサテライトリピート数を対象とした参照パネルの高精度化だけでなく、マイクロサテライト特有の変異のメカニズムを考慮したインピュテーション法の開発が必要である。これまで、マイクロサテライト周辺の単塩基変異から推定された遺伝子系図情報をもとに遺伝子系図上における変異によるマイクロサテライトリピート数の変化を考慮することで、シークエンスデータから複数検体に対してより高精度にリピート数を推定する手法を提案している（Kojima et al., Short tandem repeat number estimation from paired-end reads for multiple individuals by considering coalescent tree, BMC Genomics, 2016）。そこで、この知見をもとに参照パネルを構成する検体と SNP アレイの対象となる検体の遺伝子系図を単塩基変異情報から推定し、遺伝子系図情報を用いて SNP アレイの対象となる検体のリピート数を推定するインピュテーション法を開発する。

さらに開発手法の日本人ゲノムデータへの適用を目的とした日本人全ゲノムシークエンスデータからのマイクロサテライトにおける多型情報の整備と SNP アレイデータからの単塩基変異を主としたゲノムワイド関連解析研究を通して、関連解析において重要となる知見の収集と得られた知見を考慮した解析パイプラインの構築を行う。

３．研究の方法

本研究の方法について下記 4 項目に分けてそれぞれ述べる。

（1）シークエンスデータからのマイクロサテライトリピート数推定法開発

マイクロサテライト周辺における単位配列のリピートによる配列の特異性の低下と参照ゲノムとのリピート数の違いに起因した挿入・欠損配列によりアラインメント品質が低下することか

ら、これを補正することでシークエンスデータから高精度にマイクロサテライトリピート数を推定する手法を開発する。
（２）単塩基変異から推定された遺伝子系図情報を用いたマイクロサテライトリピート数インピュテーション法開発
遺伝子系図上での変異によるマイクロサテライトリピート数の変化を考慮し、単塩基変異から推定された遺伝子系図情報を用いてインピュテーションを行う手法を開発する。
（３）日本人のマイクロサテライトにおける多型情報の整備
東北メディカル・メガバンクプロジェクトによるゲノムコホート研究参加者のうち、1,070 人の全ゲノムシークエンスデータから常染色体におけるマイクロサテライト多型を網羅的に解析し、日本人のマイクロサテライトにおける多型情報の整備を行う。
（４）SNP アレイデータからの関連解析における知見の収集と解析パイプライン構築
東北メディカル・メガバンクプロジェクトによるゲノムコホート研究参加者のうち、約一万人の SNP アレイデータについてコホートアンケート情報等の表現型情報を対象としたゲノムワイド関連解析研究を通して、関連解析において重要となる知見の収集と得られた知見を考慮した解析パイプラインの構築を行う。

４．研究成果

マイクロサテライトリピート数をシークエンスデータから推定する手法の開発を行った。マイクロサテライトではリシークエンス解析における挿入・欠損変異などによるアラインメント品質の低下からアラインメントの誤りが多く存在するため、アラインメント結果から多くの変異パターンが観測される。このため、手法の第一段階としてリシークエンス解析結果に対して対象のマイクロサテライトにアラインメントされたシークエンスデータからマイクロサテライトの変異パターンの候補を生成する。この変異パターンの候補の生成過程において少数のシークエンスデータのみに支持されるような明らかにアラインメントの誤りによると考えられる変異パターンは除外する。第二段階として対象のマイクロサテライト周辺にアラインメントされたシークエンスデータを全て取り出し、得られた変異パターンの各候補に対して再アラインメントを行い、シークエンスデータの生成確率を計算する。計算された生成確率をもとに最も適切な変異パターンの組み合わせを選び出すことで最終的なリピート数が推定される。開発手法について Java 言語で実装を行い、シミュレーションデータと実データの双方で精度を検証している。シミュレーションデータによる精度検証では HiSeq からのシークエンスデータを模した合成シークエンスデータを生成し、複数のマイクロサテライトにおいて推定されたリピート数と真のリピート数の平均二乗誤差を指標として精度を評価している。実データによる精度検証ではNA12878検体に対するHiSeqからの公共シークエンスデータについてリピート数を推定し、長鎖型シークエンサーからの公共データを用いたリピート数推定結果を正解とした平均二乗誤差を指標として精度を評価している。精度検証において 1000 Genomes Project (http://www.internationalgenome.org/)でのマイクロサテライト多型の参照パネルの作成で使用された lobSTR と RepeatSeq を既存手法として精度比較を行っており、シミュレーションデータと実データの双方において開発手法の優位性を確認している。

明示的に遺伝子系図情報を用いると共に遺伝子系図上での変異によるリピート数変化を考慮したマイクロサテライトリピート数のインピュテーション法を開発した。開発手法では遺伝子系図上における世代間の変異によるマイクロサテライトリピート数の変化をステップワイズモデルを用いて確率的に表現している。遺伝子系図上でのリピート数の確率的な変化をグラフィカルモデルの枠組みを用いることで統計モデル化しており、これにより多数の検体のリピート数情報を確率的に同時に考慮することが可能となっている。また、遺伝子系図情報はインピュテーション対象のマイクロサテライト周辺のハプロタイプにおけるものとして単塩基変異情報から推定されている。遺伝子系図を構成するハプロタイプの中でインピュテーション対象となるハプロタイプはリピート数情報が未観測となるが、グラフィカルモデル上の Belief Propagation を用いた確率推論によりリピート数を確率的に推定することでインピュテーションが行われる。開発手法の実装は Java 言語により行っている。開発手法では遺伝子系図情報の精度が高いほど高い性能が得られることが予想される。ヒト Y 染色体では擬似常染色体領域以外において原則的にハプロタイプ間の組み換えが発生せず染色体全体の変異を用いて比較的高精度に遺伝子系図の推定が可能であるため、ヒト Y 染色体を対象として精度を検証している。具体的には、1000 Genomes Project Phase3 のデータを対象として、ヒト Y 染色体の単塩基変異情報について擬似常染色体領域や自己との組み換えが起こる可能性が高い回文領域以外に存在する単塩基変異情報から遺伝子系図情報を推定する。そして、1000 Genome Project のマイクロサテライト多型の参照パネルから得られたヒト Y 染色体の擬似常染色体領域や回文領域以外にあるマイクロサテライトリピート数情報について、一部検体の情報をマスクし推定された遺伝子系図情報を用いてマスクされた検体のリピート数を推定することでインピュテーション精度を検証している。また、Y 染色体だけでなく常染色体での精度を検証するため、ARGweaver（Rasmussen et al., Genome-wide inference of ancestral recombination graphs, PLOS Genetics, 2014）など組み換えを考慮することで染色体全体の単塩基変異から遺伝子系図の推定が可能な手法を用いて遺伝子系図情報を取得することで開発手法の精度検証を進めている。

上記手法の開発と併行して、日本人におけるマイクロサテライト多型情報を取得するため、東北メディカル・メガバンクプロジェクトによるゲノムコホート研究参加者のうち、1,070 人の全ゲノムシークエンスデータに対してマイクロサテライト変異検出手法である lobSTR を用いて網羅的なマイクロサテライトにおける多型の検出を進めてきた。このうち、法医学において広く用いられている 23 座位について、異なる日本人集団に対してキャピラリー電気泳動法により遺伝子型タイピングされた結果と比較検証を行っており、これら検証内容を含む成果について雑誌論文②として発表している。

また、東北メディカル・メガバンクプロジェクトによるゲノムコホート研究参加者のうち、約一万人の SNP アレイデータについてコホートアンケート情報から得られた肌の日焼けのしやすさと血液検査情報から得られた血中の総 IgE 値を対象とした単塩基変異を主とするゲノムワイド関連解析を行った。ゲノムワイド関連解析では、SNP アレイデータにおける各マーカーのミッシング率、Hardy-Weinberg 平衡検定の p 値、マイナーアリル頻度などによるマーカーの品質管理と各検体のコール率などによる検体の品質管理が必要となる。また、主成分分析を用いた検体の集団構造解析や検体間の近親の度合いの解析の結果を関連解析を行う際に考慮する必要がある。そこで、実際の解析を通してこれらの知見を収集し、収集した知見をもとにした関連解析パイプラインを構築している。また、肌の日焼けのしやすさとの関連解析において同定された皮膚、虹彩、髪における色素の生成に関連した変異などについて雑誌論文①として発表を行っており、血中の総 IgE 値と関連した変異ついて学会発表③で発表を行っている。

５．主な発表論文等

〔雑誌論文〕（計２件）

① Shido, K., Kojima, K., Yamasaki, K., Hozawa, A., Tamiya, G., Ogishima, S., Minegishi, N., Kawai, Y., Tanno, K., Suzuki, Y., Nagasaki, M., Aiba, S., Susceptibility loci for tanning ability in the Japanese population identified by a genome-wide association study from the Tohoku Medical Megabank Project cohort study, Journal of Investigative Dermatology, (in press) （査読有り）

② Hirata, S., Kojima, K., Misawa, K., Gervais, O., Kawai, Y., and Nagasaki, M., Population-scale whole genome sequencing identifies 271 highly polymorphic short tandem repeats from Japanese population, Heliyon, 4:1-19, 2018, doi: 10.1016/j.heliyon.2018.e00625（査読有り）

〔学会発表〕（計４件）

① 平田智士、関谷弥生、小島要、三澤計治、Gervais Olivier、人見祐基、河合洋介、徳永勝士、長﨑正朗、新たな STR8 座位のマルチプレックス PCR 法および全ゲノムシークエンスデータ解析による STR タイピング、日本 DNA 多型学会、松江市、2018 年 12 月 5 日―7 日

② 小島要、遺伝子系図情報を用いたシークエンスデータ解析と深層学習を用いた皮膚病理画像解析について、生命医薬情報学連合大会（IIBMP2018）、鶴岡市、2018 年 9 月 19 日―21 日

③ Shido, K., Kojima, K., Yamasaki, K., Gervais, O., Yen, W., Nagasaki, M., Aiba, S., A genome-wide association study identifies a novel susceptibility locus for total IgE in a Japanese population from Tohoku Medical Megabank cohort study, International Investigative Dermatology, Orland, United States, 2018 年 5 月 16 日－19 日

④ Kojima, K., Kawai, Y., Misawa, K., Mimori, T., Nagasaki, M., Comparison of short tandem repeat estimation methods with various conditions, The American Society of Human Genetics (ASHG) 2017 Annual Meeting, Orlando, United States, 2017 年 10 月 17 日―21 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：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者
研究分担者氏名：

ローマ字氏名：

所属研究機関名：

部局名：

職名：
研究者番号（8桁）：

(2)研究協力者
研究協力者氏名：
ローマ字氏名：

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。