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　　５ 月 ３１ 日現在

機関番号：８２４０１
研究種目：若手研究(B)
研究期間：2017～2018
課題番号：１７Ｋ１７５８９

研究課題名（和文）ヒトゲノムのコピー数変異における配列多様性の解明

研究課題名（英文）Sequence diversity of copy number variation in human genomes

研究代表者
三森　隆広（Mimori, Takahiro）

国立研究開発法人理化学研究所・革新知能統合研究センター・研究員

研究者番号：４０７６０１６１

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：ヒトゲノムのコピー数変異（CNV）は大規模な配列の多様性を与える変異であり、薬剤代謝酵素を含む重要な遺伝子の発現量の個人差や、様々な疾患のリスクに関係していることが知られている。次世代シークエンス技術やマイクロアレイの発展により CNV の個人差を定量できるようになってきたが、これまで配列差を含めた解析は困難であった。本研究では CNV によるゲノム多様性の解明を目指し、長鎖型シークエンサを用いた配列アセンブリ手法の改良と HLA 遺伝子での適用、短鎖型シークエンスのリード定量バイアスの低減、および SNP アレイによる CNV のインピュテーション・ジェノタイピングの融合手法を提案した。

研究成果の学術的意義や社会的意義
ヒト遺伝情報の総体であるゲノム配列は、疾患の遺伝的な要因や個人差を理解するための基礎情報であり、その解読や個人差の理解が重要である。本研究では、ゲノム配列や個人差の検出に用いる次世代シークエンサやマイクロアレイのデータを対象とし、配列の特定が難しいコピー数変異（CNV）を解読するための技術開発を行った。本研究の結果、より正確な配列の解読、コピー数定量、および未観測の遺伝子型の推定を可能にする成果が得られたため、今後配列を含む CNV の多様性が形質に及ぼす影響の理解につながると期待される。

研究成果の概要（英文）：Copy number variants (CNVs) in the human genome are mutations that give large sequence diversity, and are estimated to be present in 5-10% of the entire genome. In addition, it is well known to be associated with individual differences in expression level of important genes including drug metabolizing enzymes and the risk of various diseases. The development of next-generation sequencing technology and microarrays has made it possible to quantify individual differences in CNV, but so far it has been difficult to analyze including sequence differences. In this study, we aimed to understand the diversity of CNV including sequence differences, improved the sequence assembly method using long-read sequencer, applied it in human leukocyte antigen (HLA) region, and reduced the read quantification bias of short-read sequence data. We also proposed a fusion method of CNV imputation and genotyping with SNP arrays for genetic association studies.

研究分野：ゲノム科学

キーワード：ゲノム　コピー数変異　次世代シークエンサ　マイクロアレイ　機械学習

様 式 C－19、F－19－1、Z－19、CK－19（共通）

1．研究開始当初の背景

ヒトのような二倍体生物では、通常1細胞あたり相同なゲノム配列を2つずつ持っている。コピー数変異（CNV）はゲノムの特定領域に重複または欠失がおこることで相同配列の数が変化して起こる変異であり、個人間のゲノムの多様性の源となっている。ヒトゲノムの CNV は全ゲノムの約 5-10% の領域に存在すると見積もられており、薬物代謝酵素を含む遺伝子発現量の個人差や、精神疾患を含む様々な形質のリスクなどと関係があることが知られている。特に、コピー数の正確な定量や配列レベルでの解析は、多型の形質への影響の違いや変異メカニズムを理解するために重要である。従来、アレイ CGH や SNP アレイによるコピー数推定が行われてきたが、配列の違いを区別することは難しかった。その後現れた短鎖型・長鎖型のシークエンスデータでは、ゲノムを断片化した配列を読み取れるため、より精密な配列レベルの解析が可能になった。しかし CNV は長いリピート配列であるため、断片長以上での配列復元が難しい通常のアセンブリでは解析が困難な領域であった。この課題は、特性の異なる複数のゲノムデータや集団情報を考慮したアルゴリズムによって解決できる可能性があり、本研究で情報解析手法の開発に取り組んだ。

2．研究の目的

（1） コピー数変異（CNV）の配列多様性の解明に向けて、シークエンスデータ等のゲノム配列データを用いた高精度なコピー数解析や配列推定のための解析手法を開発することを目的とする。特に、長鎖型シークエンスデータによる配列復元、短鎖型シークエンスデータによるコピー数定量性の改善や、配列推定のための情報解析手法を開発する。
（2） 開発した情報解析手法を実データに適用し、 CNV 箇所における配列の多様性の解明を行う。これによって、将来的に変異過程の理論的なモデルの選択やパラメータの推定、特定の CNV アリルが疾患などの形質に及ぼす影響の探索や予測を可能にすることを目指す。

3．研究の方法

本研究では、ゲノムアセンブリや統計モデリングなどの要素技術を用いて長鎖型・短鎖型のシークエンスデータを解析し、 コピー数変異の状態を推定する手法の開発を計画していた。特に、長鎖型シークエンスデータによる CNV 領域を含む遺伝子領域での配列復元、短鎖型シークエンスデータを用いた配列情報に基づくリード定量補正とコピー数推定手法の開発を行った。 また、 CNV アリルと形質の関連解析研究に向け、シークエンスデータから構築した CNV のリファレンスパネルと SNP アレイデータを用いたインピュテーション解析の評価、および手法開発を行った。

4．研究成果

本研究課題において CNV 領域の配列多様性の解明に向けてゲノムデータ解析手法の開発を行い、以下の成果が得られた。

(1) 長鎖型シークエンスデータによる CNV 領域での配列アセンブリに向け、まずターゲットシークエンス用のゲノムアセンブリ手法を改善した新規手法の提案を行った。本研究課題に取り組む時点で全ゲノムデータに適用できるアセンブリ手法がいくつか提案されていたが、ほとんどが相同な染色体の配列を一本に統合したコンセンサス配列を出力するものであり、相同配列の分離は限定的にしかできていなかった。特に、研究課題とした CNV 領域ではコピー数の分離された配列アセンブリが必要となるため、 CNV 領域に適した新規の配列アセンブリ手法が必要な状況であった。一方、ゲノムの特定領域のみを読むターゲットシークエンスでは、複数の相同配列を分離するアルゴリズム（Amplicon Analysis）が開発されており、PacBio の標準解析ツール群 SMRTAnalysis で提供されている。そこで、まずこのアルゴリズムによる配列復元の検証を行った。その結果、一般にコピー数よりも多くの冗長な候補配列が出力されることや、複数のアンプリコンを共通のプライマーが増幅する場合に全長を復元できない場合があるなどの課題が分かった。本研究課題では、主にプライマー毎に分離したアセンブリや、クラスタリングによる配列統合、再アラインメントによる改善過程を入れることで、これらの問題の解決を図った。それに加えて、長鎖型の配列エラーを短鎖型シークエンスデータのコンセンサスを利用して改善するアルゴリズムを提案し、新規手法 PSARP (Primer-Separation Assembly and Refinement Pipeline) の提案を行った（図1）。

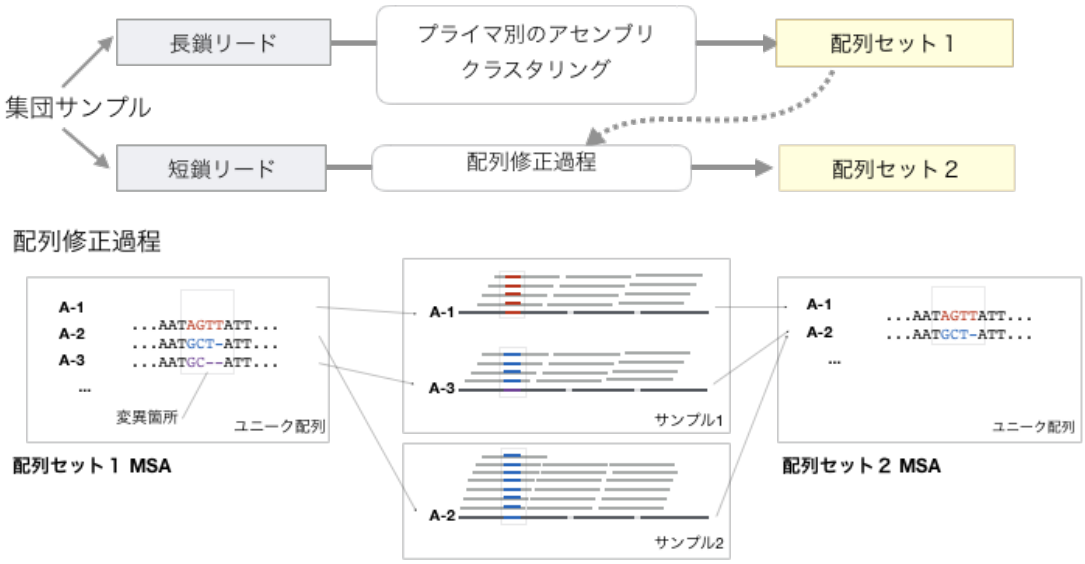

図1：長鎖リードと短鎖リードを利用し、PSARP によってターゲット領域のアセンブリと配列修正を行う過程の概要を示す。配列修正過程では、候補となる配列セットをマルチプルシークエンスアラインメント（MSA）して変異箇所を特定し、短鎖シークエンスデータのマッピング結果と比較することでアセンブリ配列の補正を行う。

(2) 前述の開発手法 PSARP は東北メディカル・メガバンク機構の日本人サンプル 208 人の class-I HLA 領域から得られた長鎖の PacBio RS II シークエンスデータに適用し、HLA-A, HLA-B, HLA-C および HLA-H を含む 139 本の HLA アリルを解読することができた。本データ解析において Amplicon Analysis の適用では多くの遺伝子で完全長のアセンブリ配列が得られなかったが、 PSARP の方法によって全ての遺伝子について完全長の配列を得ることができた。また、 PSARP では同一検体の短鎖型シークエンスデータ（HiSeq 2500）を活用し、 配列間で多様性の存在する箇所における HiSeq データとの一致率を平均約 97% から 99% まで改善することができた。また、139 本の配列を国際的な IPD-IMGT/HLA データベース（release 3.24）と比較することにより、40 本の配列が新規の多様性を含むことが分かった。既知の配列についても、全ての配列でこれまでよりも約 800 塩基 ～ 1500 塩基外側に延長した領域で配列が復元できており、既知の遺伝子型についても遺伝子の制御領域を含めた新たな多様性の存在が明らかになった。 また、偽遺伝子の HLA-H には欠失のコピー数変異が高い割合で存在するが、PSARP によるアセンブリではコピー数変異の存在しない配列も含めて高い品質でアセンブリできることを示すことができた。これらの成果について、学術論文②、および学会発表②にて発表を行った。HLA 領域は人種特異的な配列が多く存在し、組織適合性の良さや多様な疾患との関連が知られており、各人種における配列多様性の特定が重要とされているゲノム領域である。研究成果として得られた HLA 配列のデータベースは、HLA 制御領域の疾患関連解析や HLA 領域のジェノタイプ・インピュテーション、 HLA の高精度なタイピング等への活用が見込まれる。

(3). ヒト集団中のゲノム変異の多様性と形質との関連を探索する上で、多数のサンプルが得られる SNP アレイによるタイピング情報から目的とする変異を推定するジェノタイプ・インピュテーションを行う課題がある。特に、集団の持つ CNV の多様性を SNP アレイによってどの程度復元できるかを明らかにすることは重要である。そこで CNV を含むリファレンスパネルを用い、インピュテーション性能の検証を行った。まず、国際 1000 ゲノムのリファレンスパネルを用いた CNV のジェノタイプ・インピュテーションにより、 CNV の種別（欠失・重複）や長さに対する依存性を調査した。その結果、頻度の高い欠失変異については小規模な変異と同様に精度よく復元できるが、重複については長い変異ほど推定が困難になる特性があることが確認できた（図2）。一方で、 SNP アレイに含まれるプローブは長いコピー数変異になるほど増加するため、 SNP アレイ自体を用いた CNV のジェノタイプには長い変異の方が良い。これらの相補的な特性を利用したインピュテーション手法は従来知られていなかったため、本研究で新規に開発を行った。開発手法を実データに適用した結果、特にアリル頻度が 5% 以下の CNV について大幅にインピュテーション可能な変異が増えることが確認できた。本成果についてアメリカ人類遺伝学会にて発表を行った（学会発表①）。また、本研究で用いたアレイ情報の SNP 箇所のジェノタイプを利用することにより、今後配列レベルで区別したインピュテーションにも応用できることが見込まれる。

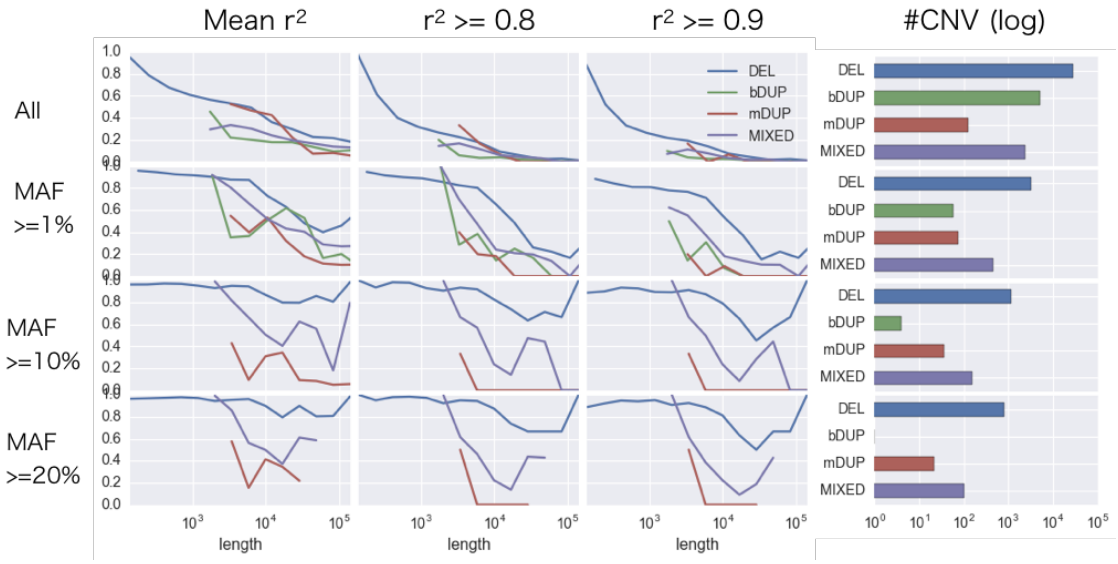

図2:国際 1000 ゲノム phase3 のリファレンスパネルを用い、イルミナのマイクロアレイ Omni2.5 の座位で Beagle 4.1 によるインピュテーション性能を計測した結果。レアな変異や長い重複変異に関してインピュテーション性能が低くなる傾向が見られる。

(4) 短鎖型のシークエンスデータによる CNV 検出とコピー数同定はリード定量によって行うため、定量のバイアスを補正することが必要になる。従来の手法では GC 含量による補正が行われているが、それに加えて GenomeSTRiP などの標準的な手法では多数のサンプルの共通トレンドによる補正も行われる。本研究ではローカス毎のバイアスを調節するオフセットと増幅率に対する正則化を考慮した手法を開発し、国際リファレンスゲノムに含まれない新規配列におけるコピー数定量解析に活用した(雑誌論文①)。一方で、ホットスポットの領域など、同一ローカスの CNV でもハプロタイプ毎に変異の境界が異なるケースも多くある。また、集団中のレアな CNV 特定という観点からも、1検体の情報のみでリード定量のバイアス補正ができることが望ましい。そこで、約 10 万塩基までのリード周辺配列を入力としてリード深度を予測する深層モデルを用い、GC 含量のみを用いた場合よりバイアスを低減させることが可能であることが分かった。今後、1検体や複数検体による CNV 検出、配列推定において上記のバイアス補正方法を活用することが見込まれる。また、同様の方法を長鎖型シークエンスデータに適用し、より大規模な CNV の配列推定に活用することが考えられる。

5. 主な発表論文等

〔雑誌論文〕(計 2 件)

① Nagasaki M, Kuroki Y, Shibata TF, Katsuoka F, Mimori T, Kawai Y, Minegishi N, Hozawa A, Kuriyama S, Suzuki Y, Kawame H, Nagami F, Takai-Igarashi T, Ogishima S, Kojima K, Misawa K, Tanabe O, Fuse N, Tanaka H, Yaegashi N, Kinoshita K, Kure S, Yasuda J, and Yamamoto M, Construction of JRG: Japanese reference genome with single molecule real-time sequencing, *Human Genome Variation*, 査読あり, accepted, 2019

② Mimori T, Yasuda J, Kuroki Y, Shibata TF, Katsuoka F, Saito S, Nariai N, Ono A, Nakai-Inagaki N, Misawa K, Tateno K, Kawai Y, Fuse N, Hozawa A, Kuriyama S, Sugawara J, Minegishi N, Suzuki K, Kinoshita K, Nagasaki M and Yamamoto M. Construction of full-length Japanese reference panel of class I HLA genes with single-molecule, real-time sequencing, *The Pharmacogenomics Journal*, 査読あり, Vol.19, 2018, 136-146, DOI: 10.1038/s41397-017-0010-4

〔学会発表〕(計 2件)

① Mimori T, Kawai T, Ueno K, Khor S, Hitomi Y, Gervais O, Tokunaga K, and Nagasaki M. Unifying copy number variant calling and imputation from SNP arrays. *The American Society of Human Genetics (ASHG) 2018 Annual Meeting*, 2018.

② Mimori T, Yasuda J, Kuroki Y, Shibata TF, Katsuoka F, Saito S, Nariai N, Ono A, Nakai-Inagaki N, Misawa K, Tateno K, Kawai Y, Fuse N, Hozawa A, Kuriyama S, Sugawara J, Minegishi N, Suzuki K, Kinoshita K, Nagasaki M and Yamamoto M. Construction of a Japanese class I HLA panel in ToMMo. 第 2 回東北メディカル・メガバンク計画合同研究会, 2017.

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。