jh190055-DAJ

大規模ゲノム情報解析にむけた数値計算技術開発と実装

徳永勝士 （国立国際医療研究センター）

概要

本研究課題では、ヒトゲノム解析において求められる大規模数値解析また電算機システム間のポータビリティを実現するために国立研究開発法人日本医療研究開発機構（AMED）で開発を進めてきたゲノム情報解析パイプラインの一部を移植すること、大規模な数値計算を行うことを目的とする。

そのため【課題 1】 数十万規模の大規模な全ゲノム情報と形質に対する RHM (Regional Heritability Mapping) による探索的パラメータ計算のためのソフトウェア改良と実装、【課題 2】「高度なゲノム情報解析基盤技術プラットフォーム」ソフトウェアの仮想環境での実行とその最適化を柱として研究開発を進めた。

その結果、利用計算機システムにおける RHM 手法の効率的なソフトウェア改良をすることで数万人規模の解析ができる体制を完了するとともに、HLA-VBSeq 改良手法等のゲノム情報解析パイプラインを当該システム上で実施できる体制を整えた。さらに、論文につながるいくつかの成果を得つつある。

1. 共同研究に関する情報

(1) 共同研究を実施した拠点名

東京

(2) 共同研究分野

□ 超大規模数値計算系応用分野

☑ 超大規模データ処理系応用分野

□ 超大容量ネットワーク技術分野

□ 超大規模情報システム関連研究分野

(3) 参加研究者の役割分担

① 全体研究推進統括

徳永 勝士

② 大規模情報解析統括

長﨑 正朗

③ 結果評価

人見 祐基

④ 情報解析

河合 洋介、植野 和子、Seik-Soon Khor、Olivier Gervais、王 妍雁、小野 彰、浅倉 章宏

⑤ 情報解析支援

男澤 良子、関谷 弥生

⑥ 数値計算評価アドバイス

関谷 勇司

⑦ スーパーコンピューター利用効率最適化

塙 敏博

2. 研究の目的と意義

人類遺伝学を専門とする研究代表の徳永勝士（国立国際医療研究センター）とバイオインフォマティクスを専門とする副代表の長﨑正朗（京都大学）はヒトゲノムの多因子疾患を対象とした高度な情報解析基盤技術プラットフォームの技術開発を目的として共同研究を行っている。

本研究では、その技術開発によって創出された一部の解析プラットフォームを「学際大規模情報基盤共同利用・共同研究拠点公募型共同研究」で提供されるプラットフォーム上で利用できるように性能評価、最適化を進めることで、学術・研究基盤の更なる高度化と恒常性を目指す。また、倫理申請および MTA において当該計算資源において利用を行うことができるデータを用いて、大規模数値解析を実施することで、現在の計算リソースのみでは実現が困難である、将来的な予防や治療に役立つ新規遺伝要因の探索を行うことを目標とする。

申請者らが開発している手法と実装を用いることで、将来的な予防や治療に役立つ多因子疾患のいくつかの新規遺伝要因を探索することができるが、現在利用可能な国立研究開発法人日本医療研究開発機構（AMED）におけるゲノム医科学用供用スーパーコンピューターで提供される計算資源には限りがある。そこで「学際大規模情報基盤共同利用・共同研究拠点公募型共同研究」で利用可能となる計算資源を用いることで、申請者らが取得した数十万人規模の UK Biobank データ（ゲノム情報及び形質データ）の解析が可能となり、またヒトゲノム中に潜む多数の新規の遺伝要因の同定を行うことで、将来的な予防や治療に役立つ成果を得ることが加速でき学術的・社会的意義が大きいと考え当研究を提案し進めた。

3. 当拠点公募型研究として実施した意義

その１）開発を進めている現在の手法と実装について、スーパーコンピューターの専門家と共同研究を行うことでより CPU およびファイルシステム双方の計算資源を効率的に利用するための実装を検討した。また、アルゴリズムの改良により、より高速に実行するための新たな手法の開発も併せて検討することができた。さらに、仮想環境（東京大学情報基盤センターの FENNEL 基盤を利用）での効率的なゲノム情報解析パイプラインの実装を進めることができた。

その２）申請者らが開発している手法と実装を用いることで、将来的な予防や治療に役立つ多因子疾患のいくつかの新規遺伝要因を探索することができる。しかし、現在利用可能な国立研究開発法人日本医療研究開発機構（AMED）におけるゲノム医科学用供用スーパーコンピューターで提供される計算資源には限りがあった。当初提案を行っていたように、研究期間中に UK Biobank の数十万人規模とその形質の利用ができる状況となった。そのため、その１の成果に併せ「学際大規模情報基盤共同利用・共同研究拠点公募型共同研究」で利用可能となる計算資源を用いることで、ヒトゲノム中に潜む多数の新規の遺伝要因の同定を進め、将来的な予防や治療に役立つ成果を得る準備体制が整えられた。

4. 前年度までに得られた研究成果の概要

本年度が初年度である。

5. 今年度の研究成果の詳細

今回ゲノム情報の「高度な情報解析基盤技術プラットフォーム」開発として、ポスト GWAS のためのいくつかの手法の開発を進めている。Regional Heritability Mapping (RHM) 手法はその１つの手法であり、微小な効果を持つ大量の遺伝子（ポリジーン）の働きに注目した手法である。同手法は、育種においてさまざまな成果がすでに得られており、また、申請者らの検証でも、いくつかのヒト疾患において新規遺伝要因の探索において成果を得つつある。しかし、ヒトゲノム解析においては、特に新しい手法であること、また、育種とは遺伝的な背景が大きく異なり、RHM で必要となるパラメータの最適な設定は現在探索的なところによるところが多い。そのためにはさまざまな疾患において探索的な計算を行うことで、適切なパラメータを検討することが求められており、ある程度のストレージおよび計算リソースを必要とするがこれらの計算資源が存在しない。

また、計算においてスーパーコンピューターに最適化されているとはいえず、計算中に出力される多数の中間ファイルの処理や、最終的な結果ファイルにまとめるための方法等の改良についても検討を行う必要がある。

これらの諸課題を解決するのが本研究の最大のポイントであり、解決することで前述の新規の遺伝要因を探索することが可能となり、ゲノム情報解析基盤プラットフォームの高度化および医療における成果を得ることができる。

そこで、以下の２つの研究課題を進めた。

(1) UK Biobank 等大規模な全ゲノム情報と形質における RHM における探索的パラメータ計算のためのソフトウェア改良と実装

(2) 仮想環境でのみ実行可能な「高度なゲノム情報解析基盤技術プラットフォーム」ソフトウェアの仮想環境での実行とその最適化

6. 今年度の進捗状況と今後の展望

前述の研究課題 (1) の予備調査として、文献 4,5,6 の成果を得ている。RHM の解析ソフトウェアについて数万人規模の遺伝型情報を効率よく保存する形式 (bgen 形式) が利用することや前処理することで効率良く入出力ができるように改良を行た。また、2019 年 12 月に UK Biobank の実データが利用できることとなったため、同データの整理を進め新しいソフトウェアが提供された電算機資源上で動作をすることを確認した。また、ジョブの投入のパターンが AMED 電算システムと異なっていたことから、ジョブの投入パターンの最適化も併せて進めた。さらに、並行して参考文献 7 の成果を得つつある。

研究課題 (2) の予備調査として、文献 1,2,3 の成果を得ていた。2019 年度に計画していたこれらの計算プログラムの移植を完了するとともに、Wang et al 2019 (文献 I) の成果をさらに発展させた研究の一部を同計算資源上で進めた (文献 II)。現在さらに発展させた解析手法を国際誌に別途投稿予定である (Wang et al in preparation)。

今後は、これらの成果を論文としてとりまとめるとともに、さまざまなゲノム情報に開発を進めたソフトウェアを適用することで新たな遺伝要因の探索を進めていく。

7. 研究業績一覧（発表予定も含む）

(1) 学術論文 （査読あり）

I. Y.Y. Wang, T. Mimori, S.S. Khor, O. Gervais, Y. Kawai, Y. Hitomi, K. Tokunaga, and M. Nagasaki. HLA-VBSeq v2: improved HLA calling accuracy with full-length Japanese class-I panel. *Hum Genome Var*, 6: 29, 2019.

II. Y.Y. Wang. Development of variational Bayesian method for the inference of major histocompatibility complex genotypes and their implementation. *Master Thesis*, 2020.

(2) 国際会議プロシーディングス （査読あり）

(3) 国際会議発表 （査読なし）

(4) 国内会議発表 （査読なし）

(5) その他（特許，プレスリリース，著書等）

(参考論文)

1) N. Nariai, K. Kojima, S. Saito, T. Mimori, Y. Sato, Y. Kawai, Y. Yamaguchi-Kabata, J. Yasuda and M. Nagasaki, HLA-VBSeq: accurate HLA typing at full resolution from whole-genome sequencing data, *BMC Genomics*, 16(2):S7, 2015.

2) M. Nagasaki *et al*, Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals, *Nature Communications*, 6:8018, 2015

3) Y Kawai, T Mimori, K Kojima, N Nariai, I Danjoh, R Saito, J Yasuda, M Yamamoto and M. Nagasaki, Japonica array: improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals, *Journal of Human Genetics* 2015; 60: 581–587, 2015.

4) T. Hasegawa, K. Kojima, Y. Kawai, K. Misawa, T. Mimori, and M. Nagasaki, AP-SKAT: highly-efficient genome-wide rare variant association test, *BMC Genomics*, 17(1):745, 2016.

5) S.-S. Khor, R. Morino, K. Nakazono, S. Kamitsuji, M. Akita, M. Kawajiri, T. Yamasaki, A. Kami, Y. Hoshi, A. Tada, K. Ishikawa, M. Hine, M. Kobayashi, N. Kurume, N. Kamatani, K. Tokunaga, and T. A. Johnson, Genome-wide association study of self-reported food reactions in Japanese identifies shrimp and peach specific loci in the HLA-DR/DQ gene region, *Sci. Rep.*, 8(1):1069, 2018.

6) S.-S. Khor, R. Morino, K. Nakazono, S. Kamitsuji, M. Akita, M. Kawajiri, T. Yamasaki, A. Kami, Y. Hoshi, A. Tada, K. Ishikawa, M. Hine, M. Kobayashi, N. Kurume, N. Kamatani, K. Tokunaga, and T. A. Johnson, Genome-wide association study of self-reported food reactions in Japanese identifies shrimp and peach specific loci in the HLA-DR/DQ gene region, *Sci. Rep.*, 8(1):1069, 2018.

7) G. Olivier, Ueno K, Kawai Y, Hitomi Y, Aiba Y, Ueta M, Nakamura M, Tokunaga K, Nagasaki M. Regional heritability mapping identifies several novel loci (STAT4, ULK4, and KCNH5) for primary biliary cholangitis in the Japanese population. *Eur J Hum Genet*, in revision.