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 １８ 日現在

機関番号：８２６２６
研究種目：基盤研究(A)（一般）
研究期間：2015～2018
課題番号：１５Ｈ０１７１７

研究課題名（和文）ミッシングヘリタビリティを埋める複合因子解析手法の開発

研究課題名（英文）Development of the combinatorial analysis methods to understand missing heritability

研究代表者
瀬々　潤（Sese, Jun）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・招聘研究員

研究者番号：４０３６１５３９

交付決定額（研究期間全体）：（直接経費）　　33,300,000円

研究成果の概要（和文）：無限次数多重検定法（LAMP）の拡張により、ミッシングヘリタビリティ(MH)の理解に繋がるソフトウエアの開発およびLAMPの拡張によるGWAS、がん体細胞変異の解析を実施した。主な成果は４つある。１．LAMPをGWASで広く利用されているPLINKに統合したLAMPLINKの開発。２．汎用GPUを利用してGWAS解析の高速化をすることで網羅的探索を可能にしたHWY開発。３．国内大規模ゲノムコホートとの連携によるLAMPを用いた解析の実施。４．長期コホート解析で利用される生存解析へ応用したSurvivalLAMPの開発である。以上を総合することで、MHの理解へと繋がるツール群が構築できた。

研究成果の学術的意義や社会的意義
遺伝性疾患の原因因子解析において、通常は単一変異と疾患の関係が調査されるが、本研究で開発したツール群を利用することで、複数の変異が同時に起こった場合に発症してしまう疾患を統計的有意性を持って特定することが可能となった。これにより、今までに比べてより多くの疾患が遺伝的に関連していることを見つけることができる可能性があるだけでなく、投薬等において、今までは副作用の起こる人と起こらない人がいる要因がわからない場合においても、遺伝的変異を用いて説明できる可能性がある。今まではあまり考えられてこなかった「組み合わせ」因子を考える重要性を本研究で示し、ツール群を整えた。

研究成果の概要（英文）：We developed softwares to understand the reason why missing heritability appears in GWAS analysis by extending Limitless-Arity Muliple-testing Procedure (LAMP). We performed four researches. 1. We developed LAMPLINK, in which LAMP is integrated into PLINK, widely used GWAS analysis software. 2. We developed HWY, in which we utilized General Purpose GPU to accelerate GWAS statistical calculations with permutation test procedure for comprehensive calculation of the statistics of pairs of SNPs efficiently. 3. We collaborated with the two largest cohorts in Japan, Biobank Japan and Tohoku Medical Megabank, and we applied LAMP to the datasets. 4. We extend LAMP to handle survival curves used in long-term cohort. It also applicable to somatic mutation analysis in cancer cells. Integrating these tools allows us to understand the keys of missing heritability.

研究分野：生命情報学

キーワード：統計的有意性　組み合わせ　GWAS　ミッシングヘリタビリティ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

ゲノム情報の取得価格が下がり世界中でゲノムワイド関連解析（Genome Wide Association Study; GWAS）が普及してきた。これによって、遺伝的疾患・表現型は何であるのか、そして、それら表現型の原因因子（原因遺伝子、原因変異）が明らかとなりつつある。一方で、特に糖尿病やリウマチ等の広く見られる疾患においては、表現型としては一定の遺伝的要因があることは認められつつも、遺伝的な因子が見つからない疾患があることも見えてきていた。これに対して、国際コンソーシアムを組んで、データを大規模に収集することで、原因因子を発見しようという研究が構想され、スタートしようとしていた。
その様な研究は、一定数の成功が得られる可能性があったが、一方で必ずしも全ての因子が発見できないことも想像に難くなかった。そのような現象は、ミッシングヘリタビリティと呼ばれていた。ミッシングヘリタビリティの要因としては、遺伝因子が発見できるほど、十分な被験者数を集められない疾患である点であること、もう一つは、当時の GWAS においては単一変異（あるいは、単一ローカス）と単一表現型の関係を見る統計学が利用されていたが、必ずしも染色体上単一の部位だけの変異が重要なのではなく、染色体全体、あるいは、複数の座位が重要である可能性が示唆されていた。このような多数の変異を扱う手法はほとんど無く、本研究ではその手法を開発し、応用することを目標とした。

２．研究の目的

本研究では、研究代表者らが開発していた無限次数多重検定法(Limitless-Arity Multiple-testing Procedure; LAMP)を拡張し、GWAS への適用をすることで ミッシングヘリタビリティの要因が一部解決できるのではないかと考えて、手法の開発・改善を実施することを目標とした。
医学・生物学においては、完全な再現実験が不可能なことも多いため、因子間の関係の発見において統計的有意性は必須のものである。また、GWAS の場合、例えば 100 万箇所の変異と特定の表現型を調査した場合、各変異と表現型を調査する検定が 100 万回実施される。ところが、変異と表現型に関連が無くとも、偶発的に関連が認められてしまう確率は、有意性の基準を 0.05、100 回の検定を考えても 1-(1-0.05)^100 = 0.994 つまり、99.4%の確率であり、100 万回の検定を実施すれば、ほぼ必ず統計的に有意な関連はなくとも、有意な関係が現れる偽陽性の問題がある。多重検定問題と呼ばれる。この問題に対し、GWAS 一般には有意水準を 5.0*10^-8 とすることで解決をしている。この水準は、前述の多重検定の問題を回避する方法として統計学で開発が進んでいる多重検定補正法の一つである Bonferroni 補正を、有意水準 0.05, 検定数を 100 万回として行った場合と同水準である。
一方で、２つ以上の変異の組み合わせを考えた場合、更に有意水準を下げる必要がある。2 個の場合であっても、100 万*(100 万-1)/2 =約 5000 億通りの計算が必要であり、GWAS の実験から得られる現実的な p 値では有意な結果が得られなくなる。３つ以上の組み合わせを考えれば、なおさらである。この問題に対し指摘した多重検定補正による有意な関係の消失を、有意な結果が得られる可能性がある組み合わせだけを考えれば良いことを示した Trone の補正と、大規模な組み合わせ計算を超高速に実施する頻出パターン解析法 LCM の二種類を組み合わせることで解決した手法が LAMP である。
LAMP は小規模なデータ・セットにおいては実際に計算することが可能であることが示されていたが、一方で理論的な問題点から大規模データに対して計算が困難であることも予想されていた。また、補正が可能であっても求めた上界が必ずしもタイトなバウンドでない可能性も高く、現実の利用は困難である可能性もあった。このため、実データへの応用を進め、問題となる点を解決していくことで、当初の目的であるミッシングヘリタビリティの解決へと向かうことが本研究の目的である。

３．研究の方法

本研究では理論・実装面で２種類、応用面で２種類の研究を並行して進めた。
理論・実装面：
第一に、GWAS において広く利用されているフォーマットに LAMP を対応させることで、広く研究成果を公表し、利用してもらえる状態を作成することを研究した。
第二に、LAMP の実行が必ずしも高速で無い場合があるため、GPGPU (グラフィックスユニットを利用した実装)の実施を研究した。
応用面：
第一に、実データの GWAS に対しての適用を考え、バイオバンクジャパン（BBJ）、東北メディカルメガバンク等との研究を推進した。
第二に、GWAS に限らずがん腫瘍における変異にも同様の問題が存在しているため、がん腫瘍の変異を対象とした研究を推進した。

４．研究成果

LAMPLINK:
LAMP は二値の組み合わせを取る手法であったが、GWAS で扱われる DNA 配列は ATGC の 4 種の塩基がありかつヒトの場合は二倍体なので父方母方の二種類を持っている、また、GWAS では一般に、父母は区別せず、対象集団で最も頻度高く現れるものを Major allele，それ以外を minor (alternative) allele として Major homo (父方母方共に major allele)，hetero (いずれかが major，もう片方が minor)などとして区別して扱われる。これらの情報を二値化した上で、GWAS の解析ができるように本研究では改良を行った。
更に、GWAS の解析で頻繁に利用されているソフトである PLINK からシームレスに LAMP が利用できるよう、PLINK に拡張機能として実装した LAMPLINK を開発[Bioinformatics 2015]し、広く配布を行った。

HWY:
GWAS の計算は変異の数、あるいは、被験者数が多くなるに従って時間を要する。複数の変異の組み合わせを考えた場合には、要する時間の増大は顕著になる。計算の高速化方法としてアルゴリズムの工夫も重要であるが、同時に近年のハードウェアの進化、特に画面を描画するために作成された GPU の計算転用による GPGPU の発展は著しく、本研究では GPGPU を用いた高速化を実施した。GPGPU による高速化は、深層学習等で用いられる行列演算には広く知られたものであったが、統計量の計算においての実績は前例がなかった。実装の結果、シングルコアの CPU に比べ、24 倍程度の高速化を達成し、GPGPU は統計量計算においても高速化可能であることを示した[ACM BCB 2015]。

国内大規模コホートでの計算：
国内の最大級ゲノムコホートであるバイオバンクジャパン、東北メディカル・メガバンクの情報を利用し、LAMPLMINK を用いたゲノム解析を実施した。計算可能性までが示せるまでに至り、現在独立コホートを用いた再現性の検証、生物学的解釈の実施を継続して行っている。

がん科学への転用：
がん腫瘍の解析においても変異解析、特に、変異の組み合わせ解析は重要である。がん科学においては通常の GWAS とは異なり、５年生存率などを表した生存曲線解析が重要となる。このような時間を経た解析は、現在の GWAS では現れていないが、今後コホートの追跡が年数を経るに従って同様に起こる問題である。このため、生存時間解析に対して LAMP を拡張した SurvivalLAMP を開発した。肺がん等の実例において、組み合わせることでより統計的に有意になる事例を検出した。

以上の様に、ミッシングヘリタビリティを埋めるための手段として、組み合わせ要因を考えた上で統計的有意性を示せる手法である LAMP の拡張の利用を提案し、手法の拡張から実際の応用まで実施した。本来であれば、これらの計算結果を生命科学的な実験で確認する必要があると考えているが、計算機的な側面が強い本提案を超え、更に科研費の年限を超える非常に時間のかかる内容となるために、本研究終了後も継続して研究を実施して、実証を継続させたい。

５．主な発表論文等

〔雑誌論文〕（計 5 件）

Raissa T. Relator, Aika Terada, Jun Sese. Identifying statistically significant combinatorial markers for survival analysis. BMC Medical Genomics. 11, 2018, 31. doi: 10.1186/s12920-018-0346-x

Aika Terada, Ryo Yamada, Koji Tsuda, Jun Sese. LAMPLINK: detection of statistically significant SNP combinations from GWAS data.Bioinformatics. 33, 2016, 3513-3515. doi: 10.1093/bioinformatics/btw418

Jun Sese, Aika Terada, Yuki Saito, Koji Tsuda. Statistically significant subgraphs for genome-wide association study. JMLR Workshop and Conference Proceedings. 47, 2015, 29-36.

Aika Terada, Hanyoung Kim, Jun Sese. High-speed Westfall-Young permutation procedure for genome-wide association studies. the 6th ACM Conference on Bioinformatics and

Computational Biology (ACM-BCB 2015). 6, 2015, 17-26
doi: 10.1145/2808719.2808721

Koichi Yamagata, Ayako Yamanishi, Chikara Kokubu, Junji Takeda, Jun Sese. COSMOS: accurate detection of somatic structural variations through asymmetric comparison between tumor and normal samples. Nucleic Acids Res. 44, 2016, e78.
doi: 10.1093/nar/gkw026

〔学会発表〕（計　4 件）

EAGLE: Explicit Alternative Genome Likelihood Evaluator. Genome Informatics Workshop 2017(国際学会). 2017 年

Identifying statistically significant combinatorial markers for survival analysis. Genome Informatics Workshop 2017 (国際学会). 2017 年

Statistical assessment for untangling higher‐order genotype‐phenotype connections. The 9th AYRCOB(招待講演)(国際学会). 2016 年 01 月 21 日. シンガポール

Statistical assessment for genome-wide association study with PPI and pathways. 生物物理学会. 2015 年 09 月 13 日金沢

〔図書〕（計　1 件）
生命情報処理における機械学習．瀬々潤、浜田道昭．講談社サイエンティフィック．2015.192 ページ

〔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
なし

(2)研究協力者
なし

※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。