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

平成　２９　年　　６　月　　２　日現在

機関番号：　８２４０１
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６８６０２２８

研究課題名（和文）バイオバンク・ジャパンサンプルに対する全ゲノムデータの構造的解析

研究課題名（英文）Structural analysis of the whole genome data from BioBank Japan

研究代表者
鎌谷　洋一郎（Kamatani, Yoichiro）

国立研究開発法人理化学研究所・統合生命医科学研究センター・チームリーダー

研究者番号：００７２０８８０

交付決定額（研究期間全体）：（直接経費）　　1,800,000円

研究成果の概要（和文）：バイオバンクジャパンデータを用いて、多因子疾患の発症に関わる遺伝因子の機械学習法による検出を試みた。Random Forest (RF)、LASSO、Elastic Net、Support Vector Machine (SVM)、Extremely Randomized Trees (ERT)、mixed Random Forest (mixed RF)の各種法を適用したところ、LASSOまたはElastic Netにより良好に検出される可能性を示した。また、LDスコア回帰法により異民族間の遺伝的因子の違いについて考察し、アトピー性皮膚炎と双極性障害に特徴的な結果を得た。

研究成果の概要（英文）：We used BioBank Japan data and applied machine learning technique to detect genetic factors associated with disease occurrence. By using Random Forest (RF), LASSO, Elastic Net, Support Vector Machine (SVM), Extremely Randomized Trees (ERT), and mixed Random Forest (mixed RF), we identified that LASSO and Elastic Net outperformed other methods. This results indicates the importance of regularization in the genetic association model.
In parallel, we applied LD score regression method or Popcorn method to see the differences of genetic background for the complex disease between Japanese and European populations from the view point of polygenic architecture. We revealed that higher heritability for atopic dermatitis and lower for bipolar disorder in Japanese compared with European population. Further studies would clarify the actual genetic regions responsible for these differences.

研究分野：　遺伝統計学

キーワード：　機械学習　ゲノム

１．研究開始当初の背景

ヒトゲノム解読以降、ゲノムワイド関連研究（Genome Wide Association Study; GWAS）が様々な疾患や量的形質の感受性遺伝的変異（特に一塩基多型 Single Nucleotide Polymorphisms; SNP）を明らかにした。GWAS は一つの疾患や量的形質に対し数十～百以上にもなる生殖細胞系列遺伝的変異が関連することを報告し、これらは今後の疾患解明のための研究に寄与するものである。一方、それらの効果量、あるいはそれらを相加的に足しあわせた効果量はとても弱いものであった。

これまでの GWAS は、一つの病気や量的形質に対して一つの SNP だけを使用した遺伝的モデルを作成し、それを全 SNP 分（50 万～1,000 万以上）繰り返す、という研究デザインであった。また、これまではまだサンプルサイズが十分大きいものではなかったため、二つ以上の SNP を用いると組み合わせごとのサンプルサイズが減少し、十分な推定を得ることは困難だった。

2014 年に理化学研究所がバイオバンクジャパン（BBJ）の全 20 万例の全ゲノム SNP ジェノタイピングを完成する予定となった。大サンプルであるため複数 SNP を使用した解析において検出力が大幅に向上することが見込まれた。

２．研究の目的

研究代表者は、ひとつのモデルに複数の SNP を取り入れ、組み合わせとして考えたときの効果を検討することによって、より強い効果量を示す遺伝的因子を同定できるのではないかと考えた。ひとつの遺伝子コード領域上に複数の変異があって組み合わさって効果が強くなるというのもなくはない可能性であるが、それ以上に、ひとつの遺伝子産物に影響するプロモーターやエンハンサー上の変異が協働したり反対に働くなどしたり、あるいは同じパスウェイ上の複数のタンパク質をコードする遺伝子の変異が組み合わせ特異的な効果を示しうるのではないかと考えることは、自然な発想であると思われた。

研究開始時点までででもそのような検討が行われてこなかったわけではないが、50 万の SNP のペアだけで考えても可能な組み合わせは 1000 億通り以上にもなってしまうため、これまでのサンプルサイズ、ならびに古典統計学的理論によってしまうと、統計学的検出力の問題から検出が極めて困難である。そこで本研究では、古典統計学手法ではなく機械学習などの新たな手法を適用することで、単一 SNP ではなく多数の SNP を組み合わせた効果を検出し、全ゲノムを構造として捉える解析を行うことを目的とした。

３．研究の方法

解析対象となるデータは、BBJ の約 20 万人の全ゲノム SNP データである。これには 47 の疾患罹患情報、ならびにアンケートまたはカルテ情報から抽出された臨床情報が付随する。

BBJ の全ゲノム SNP データは、Illumina HumanOmniExpress と HumanExome BeadChips、の組み合わせ、または Illumina HumanOmniExpressExome BeadChip により生成されたデータであり、90 万個程度の SNP データを算出する。これについて call rate < 98%のサンプルを除外したのち、call rate < 99%、Hardy-Weinberg Equilibrium test P 値 $< 1x10^{-6}$、MAF < 0.5%の SNP を各チップにおいて除外した上で全セットに存在する SNP のみからなる結合データセットを作成、さらには in-house のプログラムにより遺伝学的近縁者を除外、非東アジア人を除外した品質管理後データセットを使用した。

(1) Random Forest（RF）、Lasso、Elastic Net、Support、Vector Machine（SVM）、Extremely Randomized Trees（ERT）、mixed Random Forest（mixed RF）の各種機械学習手法を BBJ データに適用、比較した。遺伝因子による疾患表現型の説明の度合いについては、10x クロスバリデーション法により C-index を評価した。

(2)新たに GWAS による全ゲノム SNP のポリジェニック効果を評価する LD スコア回帰法（Bulik-SullivanBK et al. Nat Genet 2015; 47: 291-5.）や Popcorn 法（Brown BC et al. Am J Hum Genet 2016; 99: 1-13.）が開発されてきていたため、それらの手法を用いることで多くの SNP を同時に解析することによる遺伝的集団間の遺伝因子の違いについて考察した。日本人 GWAS 結果としては、理化学研究所の所有する GWAS 結果（SNP ごと統計量のみとなっていて個人情報が消失しているもの）を二次利用として使用した。欧州系集団 GWAS としては、公開されている GWAS 結果を取得した。

４．研究成果

(1) まず単一データで様々な機械学習手法を比較した。機械学習に入力する SNP は、GWAS の P 値が一定の閾値以下のもののみとし、その閾値を変化させて予測能の振る舞いを観察した（図 1）。すると Elastic Net または LASSO はどの SNP セットにおいても良好な予測能を示した。このいずれにおいても基本的なモデルは線形モデルとしており、これは正則化項を加えることが遺伝的モデルにおいても有用であることを示していると考えられる。一方、Support Vector Machine (SVM)の地を這うような悪い予測性能は、過剰適合してしまったことを意味しているもの

と考えられた。

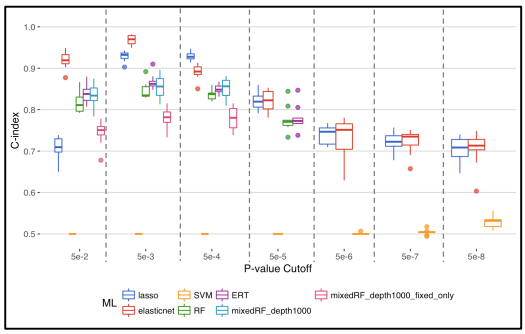

図 1 種々の機械学習手法の疾患表現型予測能と SNP セットの関係

次にここから選択した手法を用いて実際に様々な疾患を解析した。そのうち代表的なものを下記に示した。

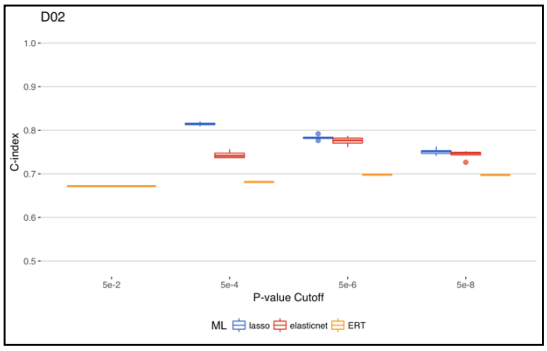

図 2 2 型糖尿病

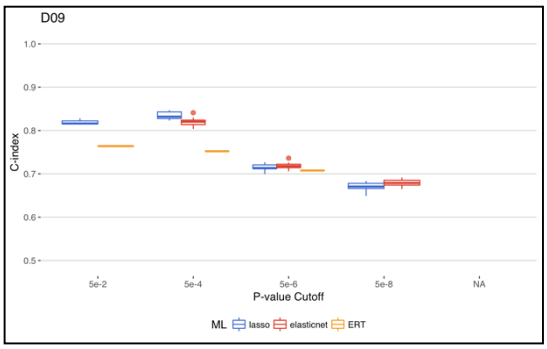

図 3 骨粗鬆症

2 型糖尿病（図 2）、骨粗鬆症（図 3）のいずれにおいても、予備的解析と同様に LASSO または Elastic Net の有用性が示された。さらに、「ゲノムワイド有意水準」として GWAS 論文報告の対象となる $P < 5x10^{-8}$ の SNP のみならず、より緩い基準で幅広く SNP を選択することで予測能が向上することを示した。

今後はこれらの結果を、同一サンプル内クロスバリデーションだけではなく、精密な前向きコホートなど独立した共同研究を募ることにより、より頑健な予測能評価を行う必要がある。

(3) LD スコア回帰法による全ゲノム SNP からの遺伝率ならびに遺伝的相関について、日本人 GWAS 結果と欧州系集団 GWAS 結果を用いて比較したところ、多くの形質については両集団で同様の遺伝率を示したが、双極性障害（欧州系で高い）とアトピー性皮膚炎（日本人で高い）において集団間の遺伝因子による貢献度の違いを認めた（図 4a）。興味深いことに遺伝的相関を見た場合（図 4b）には、民族集団間の差は消失し、遺伝的相関は生物学的メカニズムを観察していることを示唆すると考えられた。

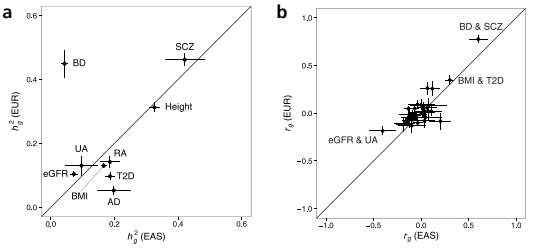

図 4 LD スコア回帰法による遺伝率・遺伝的相関推定結果。a.東アジア人集団（横軸）と欧州系集団（縦軸）の、同一形質についての遺伝率の散布図 b. 東アジア人集団（横軸）と欧州系集団（縦軸）の、複数形質の遺伝的相関の散布図。AD: アトピー性皮膚炎、BD: 双極性障害、RA: 関節リウマチ、SCZ: 統合失調症、T2D: 2 型糖尿病、UA: 尿酸値

さらに Popcorn 法により直接、異なった集団間の遺伝的相関を観察した（図 5）。

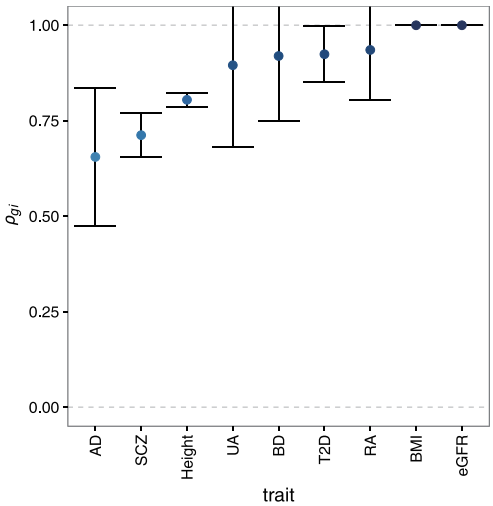

図 5 Popcorn 法による民族集団間遺伝的相関結果。略称は図 4 と同じ

アトピー性皮膚炎は相関が低く、前述の解析と一致するが、双極性障害について相関は高いと出ている。これは、アトピー性皮膚炎については実際に欧州系集団と日本人とで遺伝因子の構造が異なる可能性があるが、双極性障害については遺伝的構造は同じであるが環境因子の影響度が異なることによる結果可能

性があることを示唆する。ただし、もしこの仮説が正しいなら遺伝的相関（図 4b）においてアトピー性皮膚炎の関わる相関に差が出るとも思われる。これについては、相関を起こす要因となる共有パスウェイについては同じで、他の形質との相関に関わりない遺伝的因子のみ異なるということがありうるなら説明可能であることから、決定的な反証というわけでもない。また、いずれかの手法に理論的な問題がある可能性を否定できない。
実際に違いがあるとして、今回行った解析ではどのゲノム領域がその違いをもたらしているかを解明することは困難である。このような日本人集団と欧州系集団の遺伝因子の違いについては、さらなる考察並びに今後の研究の進展が必要である。

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　　件）

〔学会発表〕（計　1 件）
2016 年 10 月 20 日
Masahiro Kanai, Masato Akiyama, Yukinori Okada, Masashi Ikeda, Nakao Iwata, Michiaki Kubo, Yoichiro Kamatani. Trans-ethnic comparison of partitioned heritability reveals shared cell-type specific enrichment between East Asian and European GWAS. American Society of Human Genetics Annual Meeting 2016. Vancouver (Canada)
〔図書〕（計　　件）

〔産業財産権〕

○出願状況（計　　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　　件）

名称：
発明者：
権利者：
種類：
番号：
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
国立研究開発法人理化学研究所・統合生命医科学研究センター・チームリーダー
鎌谷　洋一郎（KAMATANI, Yoichiro）

研究者番号：00720880

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
国立研究開発法人産業技術総合研究所・人工知能研究センター・機械学習研究チーム・チーム長
瀬々　潤（SESE, Jun）

国立研究開発法人理化学研究所・統合生命医科学研究センター・客員研究員
小井土　大（KOIDO, Masaru）

国立研究開発法人理化学研究所・統合生命医科学研究センター・研修生
金井　正弘（KANAI, Masahiro）