様式 C-19

科学研究費補助金研究成果報告書

平成２３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：　2009～2010
課題番号：　21790342
研究課題名（和文）
Ｔｅｘｔｉｌｅ　Ｐｌｏｔによる新たな遺伝子型地図表示の提案と実装
研究課題名（英文）Design and implementation of a novel mapping technique for single-nucleotide polymorphism genotypes using Textile Plot
研究代表者
熊坂　夏彦（KUMASAKA NATSUHIKO）
独立行政法人理化学研究所・統計解析研究チーム・研究員
研究者番号：80525527

研究成果の概要（和文）：

統計学における多変量解析手法のひとつである Textile Plot を用いて，ゲノム上に存在する複数の一塩基多型を可視化し解析する新たなマッピング手法を開発し，その集団遺伝学的解釈を与えるとともに，実際のソフトウェア環境として広くインターネット上で公開をおこなった．またこの手法が疾患遺伝子解析のなかで，特にハプロタイプ解析に有効であることを，慢性Ｂ型肝炎およびクローン病・潰瘍性大腸炎のデータを用いて示した．

研究成果の概要（英文）：

Textile Plot is a novel visualization technique and multivariate data analysis tool in statistics. We applied it to single-nucleotide polymorphism genotype data to develop a novel mapping technique of linkage disequilibrium, selection footprints and disease genes. We provided a population genetic perspective of the plot along with a software environment implemented in JAVA language. We also performed disease gene mappings in an association study of chronic hepatitis B and a comparative study of Crohn's disease and ulcerative colitis in Japanese.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,700,000	510,000	2,210,000
2010 年度	1,500,000	450,000	1,950,000
年度			
年度			
年度			
総　計	3,200,000	960,000	4,160,000

研究分野：医歯薬学
科研費の分科・細目：基礎医学・人類遺伝学
キーワード：Data Visualization; 連鎖不平衡; 集団構造; JAVA; ゲノムワイド関連解析; Hardy-Weinberg 平衡; ハプロタイプ; 自然選択.

１．研究開始当初の背景

近年ゲノム医科学分野では，多種多様なゲノム情報が取得され，疾患原因の探索にはじまり個別化医療への応用も盛んにおこなわれている．特に一塩基多型(SNP)に関しては，一個人あたり約 50～100 万座位の遺伝子型を安価にしかも高速に決定できるようになったことで，世界中で疾患の感受性や薬の副作用に関するゲノムワイドな研究が盛んに行われている．また取得された SNP データの一部はネットワーク上で公開され，誰もが自由にデータを取得し，研究者独自の視点で解析を行うことも可能となっている．

その一方で，データが次々と取得され大規模化・複雑化することで，集団構造化による擬陽性や多重検定の問題など，従来の統計的手法が単純に適用できない例が多くなってきている．またデータ取得者がデータ解析をおこなうことは少なく、データ解析者は十分にデータの詳細について知らぬまま解析が進行することも多い．そこでは集団遺伝学の諸法則にそぐわない解析や統計モデルの誤用がしばしば問題となる．またデータが大規模化することで，適用できる解析方法はおのずと制限され，はじめから遺伝学的(または統計学的)なモデルを仮定した定型的解析が主流になる．それでは真の発見に結びつく柔軟な思考を妨げることになりかねない．

Textile Plot はそのような特定の解析方法やモデルに依存しない汎用な多変量解析手法として統計学の分野で提案された．Textile Plot をもちいれば，従来の解析方法やモデルにとらわれず，より柔軟にデータをながめ，そこから洞察をえることができると予想した．

実際に予備的な解析をおこなったところ，Textile Plot 上のすべての線分が水平になるような基準を用いて各遺伝子型の座標を選択することで，SNP 間の連鎖不平衡(LD)関係が視覚的にとらえられることが明らかになった．また Textile Plot では絶対 LD および完全 LD にある SNP の組は特徴的な幾何学的図形によって表現されることも明らかになった．すなわち Textile Plot では，絶対 LD は 2 SNP 間の対応するホモ接合体およびヘテロ接合体が一対一に対応し，その間の線分がすべて水平になることで表され，完全連鎖不平衡は 2 SNP 間のホモ接合体の組とヘテロ接合体の組の間に一回の交差が存在することで表わされる．通常，このような現象を定量化する指標として D' および r^2 が一般的であるが，Textile Plot をもちいれば異なるいくつかの指標を組み合わせて表示せずとも，様々な LD の度合いを包括的にとらえることができる．さらに Textile Plot では，すべてのヘテロ接合体の縦軸の位置が二つのホモ接合体のちょうど中間に位置するときに，集団がハーディ・ワインバーグ平衡(HWE)に達していることを表すという事実も明らかになり，単に SNP 間の相関を表示するだけにとどまらず，他の遺伝学的現象も反映していることが示唆された．このような現象は最初から HWE を仮定し連鎖不平衡係数を統計的に推測する HaploView などのグラフィクス表現からは窺い知ることはできない．

２．研究の目的

Textile Plot は統計学(遺伝学)的なモデルを一切仮定しないにもかかわらず，その高い汎用性から，集団遺伝学のいくつかの現象を潜在的に反映していることが予備的な研究で明らかになった．そこで本研究では，まずあらゆる SNP データに Textile Plot を適用することでゲノムワイドに遺伝子型地図を作成し，それらをもとに SNP データの背後に広がる連鎖不平衡の様子を帰納的に明らかにした．そしてその結果をもとに Textile Plot の水平性基準が連鎖平衡や HWE からの乖離をどのように表しているか数理的アプローチによって演繹的に明らかにした．さらには実際の臨床の現場で誰もが容易に利用できるようなソフトウェア環境を開発し，インターネット上で広く公開することを目指した．

３．研究の方法

本研究の遂行にあたっては，(1)実データ解析，(2)Textile Plot の集団遺伝学的解釈，そして(3)ソフトウェア開発という 3 つの局面を考えた．

実データ解析の局面では，まず国際 HapMap 計画が提供するサンプルのうち，近縁関係にない個体を含む 11 集団（西・北欧系ユタ州住民 114 人、ナイジェリアのヨルバ族 113 人、東京在住の日本人 86 人、北京在住の中国人漢民族 84 人，他 596 人）のサンプルを利用して，遺伝地図の作成を行った．また BioBank Japan に登録されたサンプルに関する SNP データを用いた日本人集団のより詳細な遺伝子型地図の作成に取り組み，疾患の有無などによる SNP のアレル頻度の違いを明らかにした．

Textile Plot の集団遺伝学的解釈を与える局面では，LD と Textile Plot の関係を明らかにするために，まず HaploView などが扱う一般的な連鎖不平衡地図と Textile Plot による遺伝子型地図との比較をおこなった．また，HWE との関係は，HWE が集団の構造化に関連していることから，Fst に代表される構造化指標との関連や，主成分分析を基本とする EIGENSTRAT との比較をおこなった．そして最終的 HWE からの乖離が Textile Plot における遺伝子型の縦軸に関する位置をどのように変化させるかを数理的に解明した．

ソフトウェア開発の局面では，上記の集団遺伝学における成果をもとに SNP データの視覚化と解析に特化したソフトウェアの開発をおこなった．具体的には，染色体の縮約図から詳細なアレルの

連鎖までを眺めるためのアルゴリズムの開発をおこない，同時にユーザフレンドリなグラフィカル・インタフェィスの実装もおこなった．プログラミング言語は OS に依存しない JAVA 言語を用い，Windows，Macintosh，Linux のすべてのプラットフォームで実行可能なソフトウェアの開発をおこなった．

４．研究成果

実データ解析の局面では，まず HapMap の SNP データを利用して遺伝子型地図を作成し LD 構造の集団間の違いを明らかにした．また集団間の違いが引き起こす HWE からの乖離に関して，ヘテロ接合体の縦軸に関する位置と HWE の関係を数理的に明らかにした．さらに Textile Plot を *LCT* 遺伝子周辺の SNP に適用することで，欧米人集団において特異的に存在する自然選択の痕跡が Long-range LD および HWE からの乖離を通して説明できることを示した(図１)．

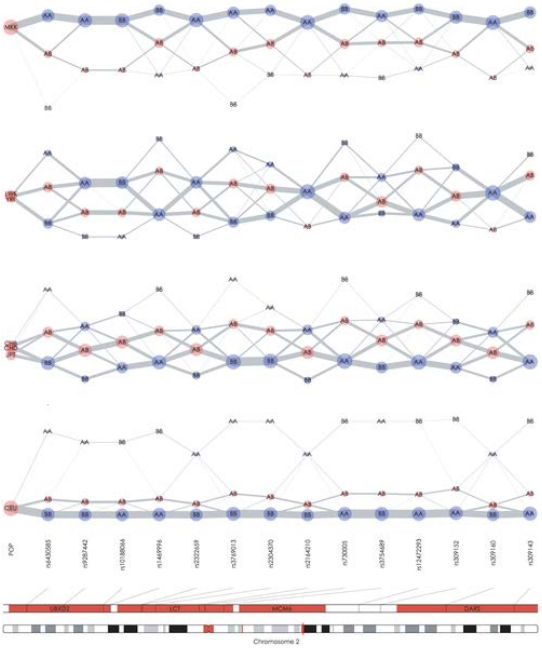

図１：HapMap データを用いた *LCT* 遺伝子周辺の Textile Plot．上段からケニア人，アフリカ系アメリカ人，アジア人，欧米人の４人種において異なる LD の様子がみてとれる．特に最下段の欧米人の Textile Plot において，ヘテロ接合体の位置が中心からずれていることから強い HWE からの乖離が伺える．また各遺伝子型を連結する線分の水平度合いからも，欧米人ではこの領域における LD が他の集団に比べ強固であることが伺える．

また BioBank Japan の SNP データを解析することで，絶対/完全 LD だけでなく，ゲノムワイドの LD 構造が視覚的にとらえることを示した．具体的には Textile Plot の縦軸の広がりが領域内の相対的な LD の強さを表していることを数学的に突き止めるとともに，実際には MHC 領域に存在する約 3,736 SNP を用いその現象を確認した．結果として 6 番染色体の 30Mbp を中心に 5’/3’側に約 4Mbp もの広がりを持つ Long-range LD の存在を視覚的に示すことができた(図２)．

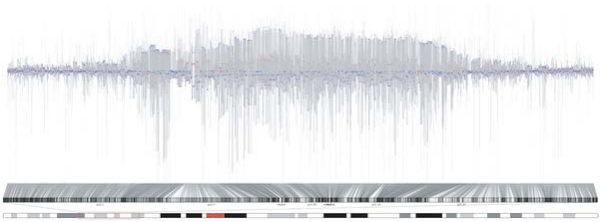

図２：BioBank 検体を用いた MHC 領域の Textile Plot．第 6 染色体 30Mbp を中心に約 8Mbp にわたって縦軸が広がっていることがこの領域の LD の強さを物語っている．

Textile Plot の疾患関連解析へ応用例としては，理化学研究所において解析された様々な疾患のうち，まず慢性 B 型肝炎に着目し，異なる HLA 型が疾患発症に与える影響を，複数の一塩基多型(SNP)の組で説明できることを Textile Plot をもちいて示した(図３)．上記の成果は全て査読付きオンラインジャーナル(PLoS ONE)に投稿し掲載されている．

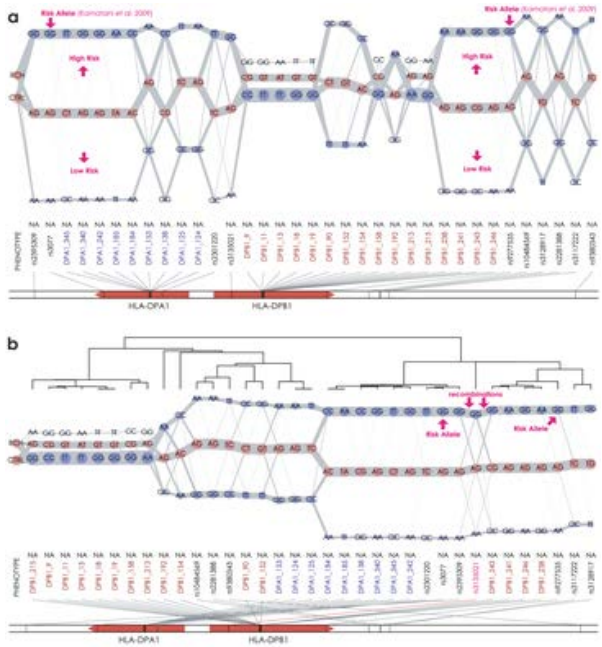

図３：BioBank 検体のうち B 型肝炎を発症した症例群と対照群の間のケースコントロール解析結果．(a)SNP マーカーを染色体上の物理位置に従って配置した Textile Plot．水平性の基準によって B 型肝炎のリスクとなるハプロタイプが図の上部に集中している様子が伺える．(b)順序付き最少距離法による SNP マーカーの並べ替え．この並べ替えによって，この領域が大きく分けて 3 つの LD ブロックからなることが明らかになった．またケースコントロール解析の結果同定された疾患感受性 SNP は同一ブロック内に含まれていることも直ちに理解することができる．

またクローン病と潰瘍性大腸炎のケースコントロール解析研究では，Textile Plot を複数アレルをもつ遺伝的マーカーに適用できるよう拡張をおこない，日本人集団において一つのハプロタイプが二つの病気のリスクになることを視覚的に示した．この成果は科学雑誌 Gastroenterology に掲載が決まった．

ソフトウェア開発の局面では，実際に JAVA 言語を用いてマルチプラットフォーム(Windows/MacOSX/Linux)での Textile Plot の実装をおこない，現在インターネット上で一般に公開されている(http://kumasakanatsuhiko.jp/projects/)．

本研究は生物学，遺伝学，数学といった既存の学問分野をこえて，SNP データに対するひとつの新しい視点を与えたことに意義があり，今後 Textile Plot の活用によって SNP を用いた疾患関連解析のさらなる発展が期待される．

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２件）

①Okada, Y., Yamazaki, K., Umeno, J., Takahashi, A., Kumasaka, N., Ashikawa, K., Aoi, T., Takazoe, M., Matsui, T., Hirano, A., Matsumoto, T., Kamatani, N., Nakamura, Y., Yamamoto, K. and Kubo, M. (2011) *HLA-Cw*1202-B*5201-DRB1*1502* haplotype confers distinct effects on ulcerative colitis and Crohn's disease in Japanese. *Gastroenterology*, doi:10.1053/gastro.2011.05.048, in press. (peer review, accepted: June 6, 2011)

②Kumasaka N, Nakamura Y & Kamatani N (2010) The Textile Plot: a new linkage disequilibrium dis-
play of multiple-single nucleotide polymorph-
ism genotype data. *PLoS ONE* 5(4)：e10207. doi:10.1371/journal.pone.0010207. (peer review, published: April 27, 2010)

〔学会発表〕（計１件）

①Kumasaka, N. and Kamatani, N. (2009) LD mapping of disease gene and haplotype analysis on textile plot, The American Society of Human Genetics 59th Annual Meeting, Oct 22, 2009, Honolulu.

〔その他〕
ホームページ等
http://kumasakanatsuhiko.jp/projects/

６．研究組織
(1)研究代表者
熊坂　夏彦（KUMASAKA N）
独立行政法人理化学研究所・統計解析研究チーム・研究員
80525527