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　３０　年　　６　月　　４　日現在

機関番号：　１２６０１
研究種目：　基盤研究(B)（一般）
研究期間：　2015～2017
課題番号：　１５Ｈ０２７７５

研究課題名（和文）階層ベイズによる高多様性領域HLAゲノムシークエンス解析法の開発

研究課題名（英文）Hierarchical Bayesian method for analyzing high polymorphic HLA genome sequence

研究代表者
井元　清哉（Imoto, Seiya）

東京大学・医科学研究所・教授

研究者番号：１０３４５０２７

交付決定額（研究期間全体）：（直接経費）　　11,800,000円

研究成果の概要（和文）：HLA遺伝子型の決定は、低いdepthが原因となりWGSからは十分な精度が得られていない。我々は、ALPHLARDと名付けたHLA型決定のためのベイズモデルを構築した。この方法は、HLAをコードする各遺伝子においてHLA型の組み合わせを6桁の解像度で決定する事が出来る。また、ALPHLARDは、データベースに登録のない新規HLA型の発見や、がん細胞のHLA遺伝子に生じた体細胞変異も同定出来る。我々は、253WES、および25WGSを用い、前者98.8%、後者98.5%の精度でHLA型を決定できることを示した。また、2,834のがんのWGSを解析し、HLA遺伝子の体細胞変異を解析した。

研究成果の概要（英文）：Although human leukocyte antigen (HLA) genotyping based on amplicon, whole exome sequence (WES), and RNA sequence data has been achieved in recent years, accurate genotyping from whole genome sequence (WGS) data remains a challenge due to the low depth. We developed a Bayesian model, called ALPHLARD, that collects reads potentially generated from HLA genes and accurately determines a pair of HLA types for each of HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, and -DRB1 genes at 6-digit resolution. Furthermore, ALPHLARD can detect rare germline variants not stored in HLA databases and call somatic mutations from paired normal and tumor sequence data. We illustrate the capability of ALPHLARD using 253 WES data and 25 WGS data from Illumina platforms and showed its high accuracy, 98.8% for WES data and 98.5% for WGS data at 4-digit resolution. We applied ALPHLARD to 2834 WGS data of PCAWG and showed somatic mutation landscape of HLA genes.

研究分野：　バイオインフォマティクス

キーワード：　ベイズモデル　HLA遺伝子型　がんゲノム　マルコフ連鎖モンテカルロ　免疫ゲノム

１．研究開始当初の背景

第 6 染色体にコードされるヒト白血球型抗原（Human Leukocyte Antigen; HLA）は、自然免疫の制御、T 細胞への抗原提示を担い、自己と非自己（細菌、ウイルスなど）を識別し、免疫反応において極めて重要な役割を担っている。多様な「非自己」に対応するために、HLA をコードするゲノム配列には、ヒトの長い歴史の中で突然変異が蓄積され、極めて多様性に富む領域となっている。HLA は、その変異の組み合わせにより、HLA 領域の A 座（HLA-A）では、例えば A*24:02:01 や A*02:07:01 のように細分化される（HLA 型とよばれる）（図１）。同様に B 座（HLA-B）、C 座（HLA-C）の型もあわせると、HLA 型の組み合わせは数万通りにものぼる。

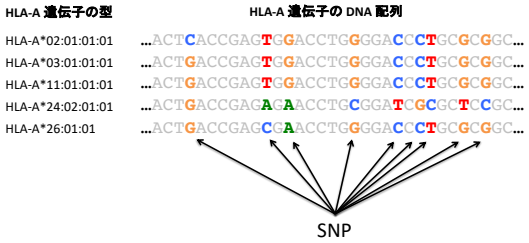

図１：SNP と HLA 型

近年のゲノムシークエンス技術の発展に伴い、全エキソームシークエンス（WES）データから HLA 型の決定（HLA タイピング）を行う研究が進められている。研究開始前年（2013 年）6 月に米国 Broad Institute とワシントン大学のグループは、ATHLATES と名付けた方法を提案した。我々は、ATHLATES の性能を評価するために、19 例の淡明細胞型腎細胞がん（ccRCC）WES データから HLA-A の型の決定を行ったところ、一定の性能を有することは確認できた(31/38=82%の精度)。一方、10 例の全ゲノムシークエンス（WGS）データ（約 30x）の解析も試みたが、ATHLATES では、2 例で片方の HLA-A 型を決めることができたのみであった。WGS データには全く対応できていないという状況であった。

また、HLA 型の決定に加えて、我々は、19 症例の腫瘍細胞の WES データを用いて HLA 領域における体細胞変異について詳しく調べた。この解析は、各症例の HLA 型を正しく決定できて始めて可能となる。結果、2 例で Loss of Heterozygosity（LOH）が起こっていた。これらの症例は、生殖細胞系ではヘテロの HLA-A 型を有していたが、腫瘍細胞では、その片方が機能を失ったものと推察される。このように、免疫学的に重要と思われる体細胞変異が HLA 領域にはみられるが、HLA 領域の多様性に正面から取り組んだ体細胞変異同定手法は、十分な研究がなされていない状況であった。

２．研究の目的

上記の状況を踏まえ、本研究課題では、生殖細胞系・腫瘍細胞の全ゲノムシークエンスデータから、HLA 型の決定を行う統計モデルの構築を行うことを目的とした。また、HLA 領域に生じた体細胞変異についても、高精度に同定を可能とする統計モデルを構築することを目的とする。構築した統計モデルを用いて国際がんゲノムコンソーシアムや The Cancer Genome Atlas（TCGA）に登録されているがんゲノムシークエンスデータを解析し、HLA 型同定を行い、HLA 遺伝子に生じ、HLA 遺伝子の抗原提示機能を喪失させる可能性のある体細胞変異のランドスケープを構築することを目的とする。

３．研究の方法

シークエンス技術の世界的な盛衰を鑑み、現在多くの研究機関において利用されている Illumina 社のシークエンサー HiSeq シリーズの生成するショートリードシークエンスデータを念頭に置き手法は開発することとした。これにより、これまでに蓄積された膨大なシークエンスデータを解析対象とすることが出来る。

まず、HLA 遺伝子からのシークエンスリードの生成モデルを構成し、どの HLA 型の配列が最も適切か選択できる方式を構築する。しかしながら、現在、データベースには登録のない未知の HLA 型が発見される可能性もあるため、サンプルの HLA 遺伝子の DNA 配列もシークエンスデータから再構築できるモデルへと拡張する。

HLA 型としては、HLA-A 遺伝子には 3000 以上、HLA-B 遺伝子には 4000 以上、HLA-C 遺伝子には 2700 以上の型が存在する。この中から例えば、HLA-A 遺伝子の遺伝子型の同定は、3000 以上の型から２つの型の組み合わせを決定するプロセスとなり、その候補数は膨大な数となり、全探索は計算時間として現実ではない。そこで、マルコフ連鎖モンテカルロ法を用いたモデルの学習を試みる。

開発する手法の精度は、日本人肝臓がん２０例と大腸がん５例の全ゲノムシークエンスデータを用いて検証する。これらのサンプルについては、サンガーシークエンスに基づき HLA 型を決定する Sequence-Based Typing（SBT）法によって HLA 型が決定されている。また、一部のサンプルについては、HLA 遺伝子領域のアンプリコンシークエスを用いてより高精度に HLA 型を決定した。SBT とアンプリコンシークエスとの結果に差異が生じた場合には、アンプリコンシークエスの結果を正しいとして手法は評価することとした。

４．研究成果

（１）まず、次世代シークエンスデータに基づき HLA 型決定を行うために提案された既存手法（ATHLETES、PHLAT）を、日本人大腸がん 5 例、肝臓がん 20 例の全ゲノムシークエンスデータを用いて、その性能を評価した。

その結果、ATHLETES は、正解率が数%、PHLAT は約５０%程度であることを確認した(Class I の A, B, C 遺伝子において､6桁同定の精度)。そこで、階層ベイズの理論に基づき、全ゲノムシークエンスデータから HLA 型を推定できる手法 ALPHLARD を開発した。手法の概略を図２にあげる。

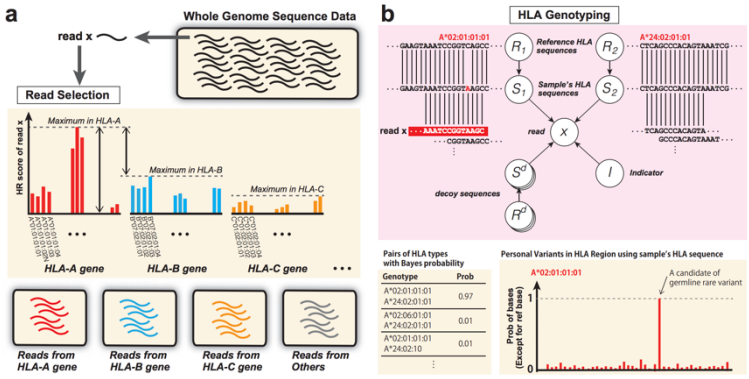

図２：構築した HLA 型決定のための手法 ALPHLARD の概念図

ALPHLARD では、読み取ったシークエンスリードを HLA の各遺伝子特異的なものかどうかを判定する（図２a)。例えば、HLA-A 遺伝子の型の決定には、HLA-A 遺伝子特異的と判定されたシークエンスリードのみが用いられる。これは、HLA-A, HLA-B, HLA-C 遺伝子がそれぞれ相同性が高く、また、HLA-A と HLA-B の間の非遺伝子領域にも相同性の高い配列が存在するため、できる限り曖昧性を排除したもので型の決定を行うための工夫である。

図２b に、構築したベイズモデルの概略を記した。x はシークエンスリードであり、HLA 遺伝子の配列(S1, S2)から生成されている。S1 と S2 は、サンプルの DNA シークエンスとなり、これはシークエンスリードから再構成するものである。R1, R2 は HLA 遺伝子の参照配列となり、これを決定することが目的となる。R1 と S1、および R2 と S2 の齟齬は許容しながらもその差が小さくなるように S1, S2 は推定され、R1 と R2 は選択される。R1, R2 の選択に関しては、マルコフ連鎖モンテカルロ法にパラレルテンパリングを組み合わせた方法を構築し、効率的に HLA 型の空間を探索できるように工夫されている。

図２b 下段左図は、R1 と R2 の組み合わせに対して事後確率が付与されて出力された例である。目安としては、トップランクに組み合わせに対する事後確率が２位のものの１０倍を超える際にユニークに決定できているというルールを用いた。図２b 下段右図は、推定されたサンプルの HLA 遺伝子の DNA 配列と参照配列の食い違いを表している。棒グラフが高いポジションの塩基は、参照配列と異なる確率が高いという事を表し、新規の HLA 型の発見に利用することが出来る。

日本人 25 名の全ゲノムシークエンスデータを用いた HLA 型決定の精度について、表１にまとめた。比較した手法は、OptiType, PHLAT, および HLA-VBSeq の３つである。HLA class I の３遺伝子、および、class II の５遺伝子の型決定を行った。表１から分かるように、我々が開発した ALPHLARD は、どの HLA 遺伝子においても最も高い精度を示した。評価した全遺伝子を通しての精度は、4-digit で 98.5% に達した。OptiType は、class I のみにしか対応していないものの 4-digit の精度は、92.6% と比較的高かった。PHLAT は、4-digit では、精度は 57.5%, HLA-VBSeq は 78.1% と実用上課題が残る結果であった。本実験を通して、我々の開発した ALPHLARD は、既存手法よりも高精度に HLA 型決定を行えることが示された。

		ALPHLARD	OptiType	PHLAT	HLA-VBSeq
HLA-A	2-digit	**100% (50/50)**	**100% (50/50)**	76.0% (38/50)	96.0% (48/50)
	4-digit	**98.0% (49/50)**	98.0% (49/50)	60.0% (30/50)	82.0% (41/50)
	6-digit	**98.0% (49/50)**	N/A	46.0% (23/50)	82.0% (41/50)
HLA-B	2-digit	**100% (48/48)**	87.5% (42/48)	72.9% (35/48)	89.6% (43/48)
	4-digit	**100% (48/48)**	85.4% (41/48)	56.3% (27/48)	75.0% (36/48)
	6-digit	**95.8% (46/48)**	N/A	39.6% (19/48)	72.9% (35/48)
HLA-C	2-digit	**100% (50/50)**	**100% (50/50)**	78.0% (39/50)	96.0% (48/50)
	4-digit	**98.0% (49/50)**	94.0% (47/50)	56.0% (28/50)	66.0% (33/50)
	6-digit	**98.0% (49/50)**	N/A	44.0% (22/50)	66.0% (33/50)
HLA-DPA1	2-digit	**100% (24/24)**	N/A	N/A	87.5% (21/24)
	4-digit	**100% (24/24)**	N/A	N/A	87.5% (21/24)
	6-digit	**100% (24/24)**	N/A	N/A	87.5% (21/24)
HLA-DPB1	2-digit	**100% (22/22)**	N/A	N/A	86.4% (19/22)
	4-digit	**100% (22/22)**	N/A	N/A	86.4% (19/22)
	6-digit	**100% (22/22)**	N/A	N/A	86.4% (19/22)
HLA-DQA1	2-digit	**100% (24/24)**	N/A	70.8% (17/24)	**100% (24/24)**
	4-digit	**95.8% (23/24)**	N/A	62.5% (15/24)	**95.8% (23/24)**
	6-digit	**95.8% (23/24)**	N/A	62.5% (15/24)	**95.8% (23/24)**
HLA-DQB1	2-digit	**100% (18/18)**	N/A	77.8% (14/18)	**100% (18/18)**
	4-digit	**94.4% (17/18)**	N/A	61.1% (11/18)	88.9% (16/18)
	6-digit	**94.4% (17/18)**	N/A	38.9% (7/18)	88.9% (16/18)
HLA-DRB1	2-digit	**100% (24/24)**	N/A	70.8% (17/24)	95.8% (23/24)
	4-digit	**100% (24/24)**	N/A	50.0% (12/24)	58.3% (14/24)
	6-digit	**100% (24/24)**	N/A	45.8% (11/24)	58.3% (14/24)
Total	2-digit	**100% (260/260)**	95.9% (142/148)	74.8% (160/214)	93.8% (244/260)
	4-digit	**98.5% (256/260)**	92.6% (137/148)	57.5% (123/214)	78.1% (203/260)
	6-digit	**97.7% (254/260)**	N/A	45.3% (97/214)	77.7% (202/260)

表１：ALHLARD の性能評価（WGS）

また、全エキソームシークエンスデータ（WES）を用いた評価についても行った。WES は WGS に比べデータの depth が厚いため、ダウンサンプリングして、データが薄くなった際にどの程度の性能低下があるかを確認した。その結果を図３に示す。

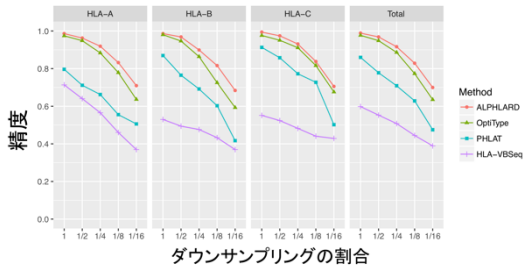

図３：ALHLARD の性能評価（WES によるダウンサンプリング）

図３の結果から分かるように、我々が開発した ALPHLARD はデータが薄い状況下でも高い精度を保持することが分かった。

この ALPHLARD を用い、正常細胞からのゲノムシークエンスから HLA 型を決定し、HLA 遺伝子の DNA 配列を予測する。また、腫瘍細胞からのゲノムシークエンスから HLA 型を決定し、HLA 遺伝子の DNA 配列を予測する。この２つの DNA 配列を比較することによりがん細胞で起こっている体細胞変異を同定することが出来る。日本人２５症例の WGS データを解析したところ、２つの一塩基置換と１つの一塩基の挿入を同定した。それらは全てサンガーシークエンスによって評価し、正しいことを確認した。

我々は、この開発した HLA 遺伝子解析の技術を、国際がんゲノムコンソーシアムと The Cancer Genome Atlas との共同プロジェクト PanCancer Analysis of Whole Genomes (PCAWG) の 2834 症例のがん全ゲノムデータに適用した。全ての症例の HLA 型を決定し、HLA 遺伝子に生じた体細胞変異を同定した。

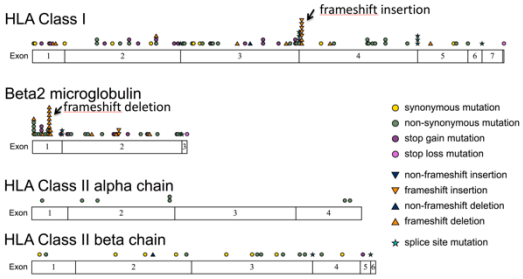

図４：PCAWG のがん 2834 症例に生じている HLA 遺伝子、B2M 遺伝子の体細胞変異

図４に結果を示す。HLA class I 遺伝子には、フレームシフトを生じる Indel のホットスポットが存在する。この変異により HLA は機能を欠失し、がん細胞が生じる変異型ペプチドを T-細胞に提示することが出来なくなると考えられる。すなわち、免疫システムによるがん細胞の排除機構が機能しなくなる可能性があり、がんの免疫回避機構の１つになっていると考えられる。

本研究により、高精度に HLA 型を決定するためのベイズモデル ALPHLARD を構築し、98%以上の精度で全ゲノムシークエンスデータから HLA 型を 4-digit で決定することに成功した。また、構築した手法を用いて、PCAWG プロジェクトにおいて活用されている 2834 症例のがん全ゲノムシークエンスデータを解析し、HLA 遺伝子に生じている体細胞変異のランドスケープを構築した。

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　３件）

1. Chapman CG, Yamaguchi R, Tamura K, Weidner J, Imoto S, Kwon J, Fang H, Yew PY, Marino SR, Miyano S, Nakamura Y, Kiyotani K., Characterization of T-cell Receptor Repertoire in Inflamed Tissues of Patients with Crohn's Disease Through Deep Sequencing, Inflammatory Bowel Dissease, 22, 2016, 1275-1285. 10.1097/MIB.0000000000000752
2. Fujii K, Miyahara Y, Harada N, Muraoka D, Komura M, Yamaguchi R, Yagita H, Nakamura J, Sugino S, Okumura S, Imoto S, Miyano S, Shiku H., Identification of an immunogenic neo-epitope encoded by mouse sarcoma using CXCR3 ligand mRNAs as sensors, OncoImmunology, 6(5): e1306617 (2017) 10.1080/2162402X.2017.1306617
3. Kiyotani Kazuma、Mai Tu H、Yamaguchi Rui、Yew Poh Yin、Kulis Mike、Orgel Kelly、Imoto Seiya、Miyano Satoru、Burks A Wesley、Nakamura Yusuke, Characterization of the B-cell receptor repertoires in peanut allergic subjects undergoing oral immunotherapy, Journal of Human Genetics, 63, 239-248, 2017. 10.1038/s10038-017-0364-0

〔学会発表〕（計　５件）

1. Shuto Hayashi, Rui Yamaguchi, Shinichi Mizuno, Mitsuhiro Komura, Satoru Miyano, Hidewaki Nakagawa, Seiya Imoto, Bayesian Method for HLA Genotyping from Whole-Genome Sequencing Data, 2015 International Workshop on Bioinformatics and Systems Biology, 2015 年 07 月 19 日~2015 年 07 月 22 日, Boston, USA
2. 井元清哉，がんゲノムビッグデータ解析から生命科学・医療へ，第９回スーパーコンピュータ「京」と創薬・医療の産学連携セミナー（招待講演），2015 年 12 月 18 日，ステーションコンファレンス東京 501
3. Seiya Imoto, Computational Methods for Cancer Immunogenomics, The 25th KOGO Annual Conference(招待講演), 2016 年 09 月 05 日～2016 年 09 月 06 日, ソウル
4. 井元清哉，がん免疫ゲノムの計算科学的解析，日本人類遺伝学会(招待講演)，2017 年
5. Seiya Imoto, Immuno-Genomic Landscape of PanCancer, The 13th ICGC Scientific Workshop(招待講演), 2017 年

〔図書〕（計　２件）

1. 井元清哉、山口類、水野晋一、中川英刀，実験医学 がん免疫療法×ゲノミクスで

変わるがん治療, 羊土社, 2017.
2. 井元清哉、長谷川嵩矩、山口類, 遺伝統計学と疾患ゲノムデータ解析(第2章-6: T 細胞受容体レパトア解析), メディカルドゥ, 2018年

〔産業財産権〕

○出願状況（計　0件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
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
井元　清哉（IMOTO, Seiya）
東京大学・医科学研究所・教授
研究者番号：10345027

(2)研究分担者
水野　晋一（MIZUNO, Shinichi）
九州大学・先端医療イノベーションセンター・特任准教授
研究者番号：40569430

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）