

機関番号：84409

研究種目：若手研究（A）

研究期間：2008～2010

課題番号：20681020

研究課題名（和文） ハプロイドDNAを用いた日本人ゲノム多様性情報基盤の高度化

研究課題名（英文） Upgrading of an information infrastructure for Japanese genome diversity using haploid DNA materials

研究代表者

久木田 洋児 (KUKITA YOJI)

地方独立行政法人大阪府立病院機構大阪府立成人病センター（研究所）・研究所・研究員

研究者番号：60372744

研究成果の概要（和文）：日本人の単一精子由来全胎状奇胎 DNA100 サンプルについて、高密度 DNA アレイを用いた SNP ジェノタイピングとコピー数多型（CNV）検出を行い、日本人ゲノムの SNP/CNV ハプロタイプを実験的に決定した。これらのデータは日本人集団における参照ハプロタイプ構造として有用である。また塩基配列多型情報を利用して同胞肺癌症例の全エクソン配列解析を行い、癌抑制遺伝子 CHEK2 上にタンパク質構造を不安定化する同症例の原因突然変異を同定した。

研究成果の概要（英文）：We determined SNP/CNV haplotype structures by analyzing 100 complete hydatidiform moles from Japanese subjects via high-density DNA arrays. The obtained haplotype is useful as a reference for the haplotype structure of Japanese population. Furthermore, using SNP/CNV information, we analyzed one lung cancer sib-pair by whole-exome sequencing, and identified a causative CHEK2 mutation that makes the functional protein unstable.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合計 |
|--------|------------|-----------|------------|
| 2008年度 | 9,200,000 | 2,760,000 | 11,960,000 |
| 2009年度 | 6,600,000 | 1,980,000 | 8,580,000 |
| 2010年度 | 3,700,000 | 1,110,000 | 4,810,000 |
| 総計 | 19,500,000 | 5,850,000 | 25,350,000 |

研究分野：複合新領域

科研費の分科・細目：ゲノム科学・基礎ゲノム科学

キーワード：ゲノム多様性、ハプロタイプ

1. 研究開始当初の背景

ヒトゲノム参照配列決定後、ヒトゲノム多様性の解明とその医学への応用がゲノム研究の中心課題となっていた。その最初の成果がハップマップ計画 (<http://www.hapmap.org/>) に集約され、次世代の多様性研究では各人類集団(人種)に関するより詳

細な解析が求められていた。西欧人集団、アフリカ人集団では各 30 家系の核家族サンプルが解析されており、家系情報を用いてほぼ正確に各 120 種類のハプロタイプが決定されていた。しかし、当時のハップマップ計画で解析された日本人のハプロタイプは集団遺伝学的推定のみによるもので、5-10%に相決

定（ハプロタイプの決定）の誤りが含まれていた。また、近親婚出生者が複数含まれ、サンプルのランダムな収集にも疑問が持たれている、等の深刻な不備があった。従って日本人サンプルの詳細なハプロタイプ構造、連鎖不平衡構造の解析は西欧人、アフリカ人と同等に行われておらず、日本人集団のさらなる解析による情報基盤整備が必要な状況にあった。しかしながら、国内外においては我々以外に日本人集団についてこれを推進しつつある研究は皆無であった。これまでに我々はハップマップ計画とは独立に日本人集団のゲノム多型情報を独自解析技術の開発とともに詳細に調べ、世界的にも非常にユニークな2種類のデータベースを構築していた。「dbQSNP」(Tahira et al. 2005)は、遺伝子転写開始点周辺領域のSNPのアレル頻度を日本人由来 DNA プールを用いたPLACE-SSCP法により高精度に決定してデータベース化したものであった。二つ目の「D-HaploDB」(Kukita et al. 2005; Higasa et al. 2007)は、ゲノムが単一精子由来である全胎状奇胎 (complete hydatidiform mole, CHM) 100 サンプルのDNAを58万個のSNPについてタイピングすることにより染色体レベルでのハプロタイプを直接決定し、データベース化したものであった。CHMの発生が稀であること(妊娠の約0.1%)とサンプルの収集が全国規模であることから、ゲノムのランダムサンプリングは十分に行われており、得られた日本人100ゲノム分の確定SNPハプロタイプ情報はハップマップ計画のそれを質量ともに凌駕するものであった。これら2つのデータベースはゲノムワイドなSNPに焦点をあわせたものであったが、近年、SNPより広いゲノム領域(数kbpから数Mbp)の挿入/欠失/重複(コピー数の多様性, copy number variation, CNV)も健常人集団に広く分布していることがアレイCGH等の実験により示されていた。当時同定されていたものだけでもヒトゲノム上の20%弱の領域を占め、疾患原因候補領域と一致している場合も多かった。このことは、健常人集団中に潜在的な疾患候補者が存在することを示していた。それでCNVと表現型の関連(量的効果等)を解明していくための一次情報として、CNV領域とSNPなどの多型との正確で詳細なハプロタイプ情報を整備する必要があったが、健常人組織由来のゲノムDNA(2倍体)を使った解析では家系情報無しに確信を持ってハプロタイプを決定することは困難であった。そこで、我々が保有していた日本人CHM由来ハプロイドDNAを利用することで、これまでより高精細なSNPとCNVを含むハプロタイプ構造を決定し、次世代日本人集団確定ハプロタイプマップを作製することを計画した。

2. 研究の目的

日本人集団多様性情報の医学への応用を目指した情報基盤整備を行うために、日本人ゲノムの高精度SNP/CNVハプロタイプマップを作成し、それをデータベース化することを目的とした。具体的には、次のことを行う。

1) 日本人由来CHMゲノム100サンプルについて、高密度DNAアレイを使い、約百万個のSNPについてジェノタイプを決定する。また、アレイ上のプローブシグナル強度情報を利用し、ゲノム上のCNVを検出する。

2) SNP及びCNV情報を統合解析し、我々が以前作成していた日本人のSNPハプロタイプマップを大幅に発展させた高精細SNP/CNVハプロタイプマップを作製し、これに基く次世代日本人ゲノム確定ハプロタイプデータベースを構築し、医学分野での応用に向けた多型情報基盤を整備する。

本研究はポストゲノム時代の疾患研究の最も根幹を成す研究であるため、日本人集団に特化した疾患に特徴的な遺伝子変異の探索や新規がん進行度診断方法の開発などに役立つものである。

3. 研究の方法

(1) 全胎状奇胎(CHM)ゲノムを用いた全ゲノムSNP/CNVハプロタイプの直接決定
CHM100サンプルから抽出したゲノムDNA試料について、Affymetrix社及びIllumina社の高密度オリゴヌクレオチドアレイ(SNP6.0, Human1M)によりゲノムワイドなSNPのジェノタイプデータを収集した。CNV領域の検出はアレイ実験の膨大なシグナル強度データをもとに行った。検出した一部のCNVについてはリアルタイムPCR法によりコピー数の増減確認実験を行った。また、各CHMで検出されたCNVを領域ごとの重なり具合からCNVイベント(CNVE)、CNV領域(CNVR)にまとめ、それらと周辺のSNPハプロタイプ構造との関連を調べた。

(2) 日本人ゲノム確定ハプロタイプデータベースの構築
実験的に直接決定した日本人由来CHMゲノムの各SNP/CNVハプロタイプマップをデータベース化した。ゲノムブラウザGeneric Genome Browserを使い、SNPジェノタイプデータをWeb上で公開した。

(3) 塩基配列多型情報と次世代シーケンサーを利用した全エクソン解析による同胞がん症例の原因遺伝子異常の探索
既存のマッピングソフトウェア、変異検出ソフトウェア及び自作のperlスクリプトを組

み合わせるにより、次世代シーケンサーによる疾患原因異常を効率よく解析する為の個人ゲノムデータ解析システムを構築した。またこのシステムの中には、本研究及び公共データベースの多型情報を利用して、既知及び新規変異を区別する機能を組み込んでいる。テストケースとして、大阪府立成人病センターで収集された同胞肺癌症例の血液からゲノム DNA を抽出し、Agilent 社 SureSelect を用いたエキソン濃縮及び次世代シーケンサーによる全エキソン配列解析 (whole-exome sequencing) を行った。また、ゲノムワイドなゲノム構造変化の有無を調べるために、Agilent 社 HumanGenome244K を使ったアレイ CGH を行った。全エキソン配列解析による変異検出の精度を評価する為に Illumina 社 Human1M アレイによるゲノムワイドな SNP タイピングデータを得た。さらに SNP タイピングデータについては、PLINK (<http://pngu.mgh.harvard.edu/~purcell/plink/>) を使用して、同胞のホモ接合マッピング解析を行った。

4. 研究成果

(1) 全胎状奇胎 (CHM) ゲノムを用いた全ゲノム SNP/CNV ハプロタイプの直接決定
CNV を検出する際、2 倍体ゲノムを用いるより、ゲノム全体がホモ接合状態である CHM の方が、コピー数の変化状態が大きいため高感度に検出されることが考えられる。Affymetrix 社 SNP6.0 アレイと Illumina 社 Human1M アレイでの CHM サンプル (一部) と通常の 2 倍体サンプル (CHM 患者の血液由来) の解析結果を比較すると、CHM ではるかに多くの CNV が検出されやすいことを確認した。引き続き、100 個の CHM について 90 万個の SNP マーカー及び 90 万個のコピー数解析用マーカーを含む SNP6.0 アレイで SNP タイピング及び CNV 検出を行った。平均して 1 ハプロイドゲノム当たり 80 個、領域としてゲノムの 3.1 Mb を占める CNV が検出された。この CNV 解析結果を取り入れ、SNP ジェノタイプ判定を厳密に行うことにより (具体的には、欠失領域の SNP ジェノタイプ判定をしない)、不確実性を排除した SNP ジェノタイプデータを得た。また、マーカーシグナルのプロファイルや SNP 判定率を基に DNA の品質に問題があった CHM サンプルを除くことにより、精査された 85 個の CHM の SNP/CNV ハプロタイプを得た。CHM サンプル間でゲノム上の位置が重なる CNV は、1336 の領域にまとめることができた (CNVR)。それらの内、サンプル間で良く見られる (頻度 5%以上) CNVR について、周辺の SNP アレルとの関連を調べてみると、約半数は連鎖不平衡の状態にあった ($r^2 > 0.8$)。さらに、CNVR に含まれる CNV を

重なり具合 51%を基準にしてまとまりを細分化した場合 (CNV イベント, CNVE) で調べると、特定の SNP ハプロタイプ上では長さの異なる CNV が頻繁に生じやすいことを示唆するデータを得た。これは SNP がほぼランダムに生じているのとは対照的である。このことから、一般集団に観察される CNV の形成の他、癌の進行過程で見られるゲノム構造変化 (体細胞変異) にも特定のハプロタイプ構造が関与していることが考えられる。

またハプロタイプマップのマーカー密度をさらに高めるために、100 万 SNP マーカーを搭載した Illumina 社 Human1M アレイを用いたタイピングデータ (100 サンプル分) を得た後、SNP アレイ 6.0 での解析結果と統合した。28 万個の SNP は両アレイで重複するので、それらを使って両アレイでのジェノタイプピング一致率を調べると、99.9%以上という極めて高精度な結果であった。よって両解析を合わせることでヒトゲノム 2 kb ごとに 1 SNP がタイピングされた高密度 SNP ハプロタイプマップを得ている。

(2) 日本人ゲノム確定ハプロタイプデータベースの構築

実験による CHM ゲノムの解析から得られた日本人 SNP/CNV ハプロタイプデータをデータベース化し、Generic Genome Browser を使ったブラウザで、遺伝子情報と共にタイピングデータを表示する D-HaploDB を構築した。本研究期間内に SNP6.0 アレイでの解析結果を基にしたバージョン 3 (D3) を web 上で公開している。そのアップデート版となる Human1M アレイを追加したもの (D4) については、研究期間内にデータベース化まで終了しており、今後公開していく。これらのデータは日本人集団における様々な疾患や体質の遺伝的背景を解明する為の重要な情報基盤となる。

(3) 塩基配列多型情報と次世代シーケンサーを利用した全エキソン解析による同胞がん症例の原因遺伝子異常の探索

次世代シーケンサーによる患者個人のゲノム解析では多くの変異が検出されるので、候補疾患原因変異と多型を区別することが不可欠である。本研究により得られた日本人集団でのハプロタイプ及び多型情報は公共の多型データとともに既知及び新規変異を区別する為に有用な情報源である。今回構築した次世代シーケンサーを利用する個人ゲノムデータ解析システムの運用テストとして、特徴的な症状を呈する同胞肺癌症例の全エキソン配列解析を行った。全遺伝子のタンパク質コード領域に対し、約 100 倍の冗長性を塩基配列決定を行い、ヒトゲノム参照配列と比較して、各患者での変異部位を特定した。

既知多型を排除し、アミノ酸置換を伴う新規変異を各患者について約 500 個ずつ検出した。さらにゲノムワイドな SNP アレイ解析により、1 Mb 以上のホモ接合領域が多数存在することを発見し（患者あたり 200 Mb）、本症例が血族結婚を経ている事を明らかにした。シーケンス結果と SNP アレイによるホモ接合マッピング解析を統合することにより、22 番染色体上の癌抑制遺伝子 CHEK2 に同遺伝子産物を不安定化させるホモ接合突然変異を同定した。この遺伝子の不活化が肺癌の発症に強くかかわっていると考えられ、今後この成果を新規肺癌マーカーの開発に役立てていくことを予定している。

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 6 件)

- ① Kawarazaki S, Taniguchi K, Shirahata M, Kukita Y, Kanemoto M, Mikuni N, Hashimoto N, Miyamoto S, Takahashi JA, Kato K. Conversion of a molecular classifier obtained by gene expression profiling into a classifier based on real-time PCR: a prognosis predictor for gliomas. *BMC Medical Genomics*, 3:52, 2010. 査読有
- ② Kukita Y, Yahara K, Tahira T, Higasa K, Sonoda M, Yamamoto K, Kato K, Wake N, Hayashi K. A definitive haplotype map as determined by genotyping duplicated haploid genomes finds a predominant haplotype preference at copy-number variation events. *American Journal of Human Genetics*, 86:918-928, 2010. 査読有
- ③ Umeno J, Matsumoto T, Esaki M, Kukita Y, Tahira T, Yanaru-Fujisawa R, Nakamura S, Arima H, Hirahashi M, Hayashi K, Iida M. Impact of group IVA cytosolic phospholipase A2 gene polymorphisms on phenotypic features of patients with familial adenomatous polyposis. *International Journal of Colorectal Disease*, 25:293-301, 2010. 査読有
- ④ Higasa K, Kukita Y, Kato K, Wake N, Tahira T, Hayashi K. Evaluation of haplotype inference using definitive haplotype data obtained from complete hydatidiform moles, and its

significance for the analyses of positively selected regions. *PLoS Genetics*, 5:e1000468, 2009. 査読有

- ⑤ Kuga D, Mizoguchi M, Guan Y, Hata N, Yoshimoto K, Shono T, Suzuki SO, Kukita Y, Tahira T, Nagata S, Sasaki T, Hayashi K. Prevalence of copy-number neutral LOH in glioblastomas revealed by genomewide analysis of laser-microdissected tissues. *Neuro Oncology*, 10:995-1003, 2008. 査読有
- ⑥ Guan Y, Hata N, Kuga D, Yoshimoto K, Mizoguchi M, Shono T, Suzuki SO, Tahira T, Kukita Y, Higasa K, Yokoyama N, Nagata S, Iwaki T, Sasaki T, Hayashi K. Narrowing of the regions of allelic losses of chromosome 1p36 in meningioma tissues by an improved SSCP analysis. *International Journal of Cancer*, 122:1820-1826, 2008. 査読有

[学会発表] (計 7 件)

- ① 久木田洋児 全エキソン配列解析による同胞がん罹患症例の原因遺伝子変異の探索第 3 3 回日本分子生物学会年会第 8 3 回日本生化学学会合同大会, 2010 年 12 月 7 日, 神戸
- ② Kukita Y. Detection of human copy number variations using complete hydatidiform moles carrying haploid genomes. 58st Annual Meeting of the American Society of Human Genetics, 2008 年 11 月 14 日, 米国・フィラデルフィア
- ③ Kukita Y. Detection of human copy number variations using a collection of complete hydatidiform moles: a strategy of CNV-detection using human haploid genomes. Cold Spring Harbor Laboratory Meeting "The Biology of Genomes", 2008 年 5 月 8 日, 米国・ニューヨーク

[図書] (計 1 件)

- ① Tahira T, Kukita Y, Higasa K, Okazaki Y, Yoshinaga A, Hayashi K. "Estimation of SNP allele frequencies by SSCP analysis of pooled DNA" in *Single Nucleotide Polymorphisms*, Komar AA ed. *Methods in Molecular Biology*. Humana Press Inc. pp193-207, 2009.

[その他]

データベース

D-HaploDB (全胎状奇胎ゲノム DNA を用いて直接決定した日本人集団ゲノムのハプロタイプのデータベース, <http://orca.gen.kyushu-u.ac.jp>)

6. 研究組織

(1) 研究代表者

久木田 洋児 (KUKITA YOJI)

地方独立行政法人大阪府立病院機構大阪府立成人病センター (研究所)・研究所・研究員

研究者番号：60372744