科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月　 ２ 日現在

機関番号：１７１０２
研究種目：挑戦的萌芽研究
研究期間：2012～2013
課題番号：２４６５７００５

研究課題名（和文）日本人ハプロイドゲノムの超並列シークエンス解析による構造多型の完全解明

研究課題名（英文）Elucidation of structural variations in Japanese genome by massively parallel sequencing of haploid genome

研究代表者
田平　知子（Tahira, Tomoko）

九州大学・生体防御医学研究所・講師

研究者番号：５０１５５２３０

交付決定額（研究期間全体）：（直接経費）　　3,100,000円、（間接経費）　　930,000円

研究成果の概要（和文）：ヒトゲノムの構造多型は遺伝情報の個人差に大きく寄与しているが、通常の２倍体細胞のシークエンス解析によってその構造を決定することは困難である。全胞状奇胎は単一精子由来の倍加ハプロイドゲノムで全領域ホモの多型情報が得られる。この利点を生かし、超並列シークエンスにより全ゲノム新規アセンブリを行うことを検討した。しかし、現時点の精度および読み取り長では実現が困難であることが判明した。そこで、マイクロアレイ解析により検出された構造多型を個別に解析する方針とした。主に薬物代謝関連遺伝子を解析し、GSTA1-GSTA2遺伝子間の非アレル間相同組換えによる欠失が日本人集団に低頻度にあることを新たに見出した。

研究成果の概要（英文）：Structural variants (SVs) account for significant portion of genomic variability, but still remain difficult to map. Delineating SVs from sequence reads of diploid cells are difficult, because most of the SVs are heterozygous, and defining the haplotypes of overlapped SV regions directly from the read data are inherently unsolvable. Genomes of complete hydatidiform moles (CHMs) derived from single sperms are genome-widely homozygous and can provide definitive haplotype information of SVs. Our initial plan was to assemble whole human genome de novo by sequencing several CHMs. However, it turned out that paralogous sequences cannot be precisely mapped even by current technology of massively parallel sequencing. Thus we changed our focus to define breakpoints of SVs detected by microarray analysis of 84 CHM samples. We studied SVs in pharmacogenes and identified new deletion between GSTA1 and GSTA2 that produced a hybrid gene by non-allelic homologous recombination.

研究分野：生物学

科研費の分科・細目：基礎生物学　遺伝・ゲノム動態

キーワード：ヒトゲノム構造多様性　ハプロタイプ

１．研究開始当初の背景

シークエンス解析の超高速化・技術革新によりヒト全ゲノム配列決定を短期間で行うことが可能になった．しかし，通常の２倍体細胞由来のディプロイドゲノムの解析ではヘテロ部位での配列の解釈の間違いが起きやすく，その影響を少なくするため多数回同じ配列を読み取る必要がある．また，コピー数多様性（CNV）のようなゲノム中の複雑な構造多型をディプロイドゲノムの解析によって決定することは，ヘテロ部位での配列の解釈が困難であるため不可能に近かった．われわれは全胞状奇胎（complete hydatidiform mole:以下 CHM と略す）が１精子由来のゲノムが倍加したハプロイドゲノムを持つことに着目し，日本人由来の CHM のゲノムを種々のマイクロアレイによりタイピングし，その結果を確定的ハプロタイプデータベース（D-Haplo データベース）として公開してきた．CHM ゲノムは全領域ホモ接合であるため，CNV が感度よく検出される．84 検体について Affymetrix SNP 6.0 および Illumina 1M-Duo の２種類のゲノムワイドマイクロアレイでタイピングを行い，170 万個の SNP 配列を決定し，CNV に関しては，いずれかの検体で検出されたものをマージすることにより 2339 領域(CNVR)を検出した（D-Haplo データベース Phase４として公開）．しかし，マイクロアレイによる CNV 解析ではマーカーの密度と定量可能範囲に限界があることから，構造多型も含むハプロタイプ決定には全ゲノムシークエンス解析が必要であると考えた．

２．研究の目的

ヒトゲノム参照配列は，ゲノムワイドな遺伝子発現解析や疾患関連解析の基盤となるものであるが，同配列（研究開始時は NCBI Build37）は複数の個人（ディプロイドである）から得られた配列断片をつなぎ合わせたものであり，複雑な構造多型を持つ領域では実際の染色体におけるハプロタイプを反映していないという問題があった．また，参照配列に含まれていない配列も多数あることが知られていた．これらの問題点を解決するためには単一精子由来ハプロイドゲノムをもつ全胞状奇胎試料を直接シークエンスし，de novo assembly を行うのが最良の方法であると考えた．

このような「ハプロイドゲノムを利用したハプロタイプ直接決定」は遺伝子重複から生じた複雑な領域の配列解析に有用であることは明白であり，実際にヒト参照配列コンソーシアム（GRC）でも，高度に構造多型があり解析が困難な領域では全胞状奇胎由来の１つの細胞株のゲノム解析を行っている．われわれは，80 検体以上の全胞状奇胎 DNA を有し，そのゲノムワイドなジェノタイプデータを取得していることより，これを日本人の参照ゲノム配列の構築に役立てたいと考えている．

本研究では超並列シークエンサーを用いてこのハプロイドゲノムのリソースの全配列を塩基配列レベルで直接決定し，次世代のヒト参照配列として提供するための基礎研究を行い，その実現性を検討することを目的とする．

３．研究の方法

(1)構造多型の情報学的解析

1000 ゲノムプロジェクト phase I で検出された構造多型の情報については EBI のサイト（ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/consensus_call_sets/sv/）から取得した．また DGV および GENCODE version 17 の情報は UCSC のサイト(http://genome.ucsc.edu/)から Table Browser を用いて取得した．BedTools（Quinlan & Hall，Bioinformatics 2010）により領域のオーバーラップを検討した．

(2) 構造多型の切断点の塩基配列決定

全胞状奇胎 DNA あるいはそれを全ゲノム増幅したものを鋳型として切断点(breakpoint)周辺領域を PCR により増幅し，PCR 産物の塩基配列決定をサンガー法によりキャピラリーシークエンサーを用いて行った．

４．研究成果

本研究は，当初の計画では全ゲノムを超並列シークエンサーで解析し新規にアセンブルすることで構造多型を確定することを目指していた．しかし，構造多型の多くが配列重複領域にあり，その解析のためには高精度・長鎖リードが不可欠である．現状での全ゲノム解析の技術でのショートリードのマッピングには限界があり（Lee & Schats Bioinformatics 28，2097-2105，2012），目的を達成するのは困難と想定された．また，ショートリードで読み取り配列を修正した長鎖リードで新規にアセンブル行う「hybrid error correction」（Koren et al. Nature Biotech. 30，693-700，2012)は有効であると想定されたが，コスト的に不可能である．そこで，D-HaploDB Phase4 で検出した CNV 領域(D4-CNVR)に絞って塩基配列解析を行い切断点を決定することを考えた．

ここで，まず D4-CNVR について情報学的解析によりその特徴を検討した．CNV の網羅的なデータベースである DGV（July 23，2013）に登録されている約 20 万個の CNV とのオーバーラップを検討したところ,D4-CNVR の 95%以上が DGV に登録されている CNV と１塩基以上重なっており，サイズが 10 kb 以上のものでは７割以上が 50%以上の重なりによりオーバーラップしていることが分かった．

次に 1000 ゲノムプロジェクトで主として深度の浅いシークエンスにより同定された SV (biallelic deletion のみ)と比較したところ，サイズが 10 kb 以上のものは半数以上が 50%以上の重なりによりオーバーラップしていた（図１）．マイクロアレイ解析により切断点を予測することは難しいが，10 kb 以上の欠失として検出されたものは真の切断点に近いと予想された．

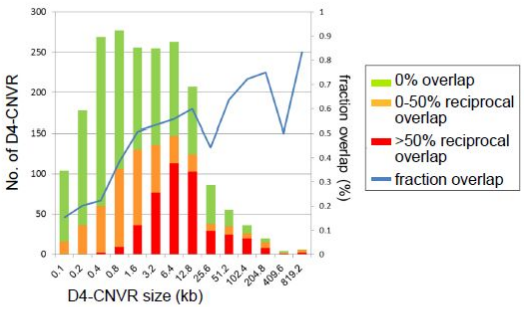

図１　1000 ゲノムプロジェクト（Phase1）で同定された SV(946 検体から同定された 14,422 個の欠失)との比較

さらに D4-CNVR とタンパク質をコードする遺伝子のエクソンとのオーバーラップを検討した．表 I に示すように，245 個の D4-CNVR がエクソンとオーバーラップした．ここでの“Loss”，“Gain”は相対的なもので，必ずしも欠失あるいは獲得を意味していないが，エクソン領域の（おそらく生体にとって不利な）“Loss”は抑制されていると考えられる．

表１　エクソンとのオーバーラップ

	Total	Loss	Gain	Both
D4-CNVR	2339	2016	271	52
Overlap with exon(s)	245	143	71	31

D4-CNVR のなかで最も頻度が高かったのは UDP-グルクロノシルトランスフェラーゼをコードする UGT2B17 遺伝子領域の欠失であった（88%）．そこで薬物代謝・薬物応答に関連する遺伝子に着目し，“pharmacogene”と分類されている 253 遺伝子（Gamazon et al. Pharmacogenet Genomics. 22:261-267, 2012）のなかで D4-CNVR とオーバーラップする 8 遺伝子を抽出した．そのうち，1 検体のみで検出された CYP4F2，SULT1A1，およびテロメアに近い領域で“Gain”として検出された CYP2E1 を除外して，残る UGT2B17，GSTT1，GSTM1，CYP2A6，GSTA1 の領域の構造多型を解析した．これらはすべて非常によく似た配列の間で NAHR (non-allelic homologous recombination，非アレル間相同組換え)が起きることにより形成されたと考えられる．

UGT2B17，GSTT1 については欠失の切断点が

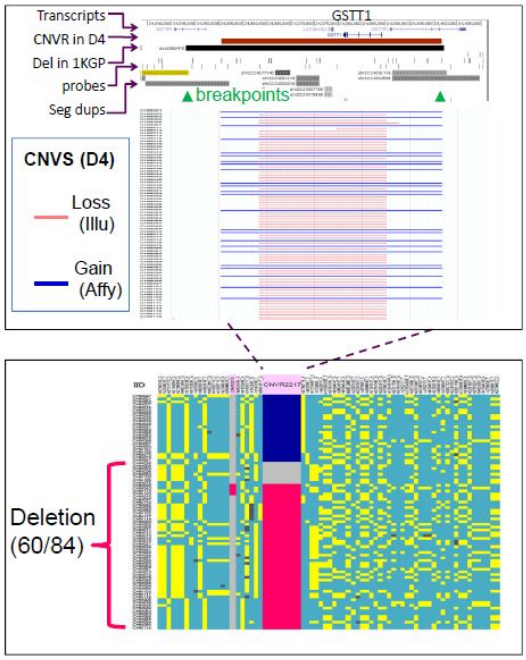

図 2 GSTT1 領域の欠失
上段は D-HaploDB での“Loss”(赤い横線が各サンプルにおける領域，つまり CNV セグメントである)と“Gain”(青い横線)．実際の切断点は外側にある(緑三角)．シークエンス解析の結果，青で示したサンプル以外が欠失であることが判明した．ここで，青は Affymetrix のアレイのコピー数を定量する際にレファレンスとして全サンプルの平均値（2 コピー以下）を用いたために見かけ上“Gain”となっているが，実際は正常コピー数と考えられる．下段は CNVR と周辺の SNP からなるハプロタイプを示している．

文献的に報告されており，CHM 84 検体のシークエンス解析でも同じ部位で欠失が検出された．GSTT1 はグルタチオン S-転移酵素 θ をコードしているが 71%のサンプルで欠失していた（図 2）．また，この欠失と最も強く連鎖していたのは rs5760176 (r^2=0.83)であったが，この SNP のジェノタイプが GSTT1 の発現と関連していることが報告されている

GSTM1 については，切断点の決定は成功していない．CYP2A6-CYP2A7 領域の欠失は 1000 ゲノム計画では 2 種類あることが報告されており，またこの領域の多型はこれまで詳細に調べられている．この領域についてハプロタイプ決定を次世代シークエンサーにより決定しようと考え，ロング PCR により欠失型および遺伝子融合型の 2 種類の PCR 産物を得ているが，まだ解析には至っていない．

これらのように頻度が高い既知の欠失に加えて，新たに GSTA1-GSTA2 領域の低頻度の欠失も検出した．図 3 にその領域の構造を示す．PCR 産物（A,B,C）のシークエンス解析によりこの領域で長い欠失と短い欠失がオーバーラップしていることが確認された．長い欠失の切断点は 1000 ゲノムプロジェクトで

検出されているものと異なり，GSTA1 および GSTA2 のイントロン 2 にあり，相同な領域での組換えにより融合遺伝子が形成されていることが判明した．これにより，コードするグルタチオン S-転移酵素αのアミノ酸配列の一部が変化することになり，また発現量が減少することにより，薬物代謝に影響を及ぼすことが予想される．同じ欠失はヒトリンパ芽球細胞でも検出された．この結果は CHM ゲノムの解析が低頻度の構造多型を解明するのに有用であることを示している．

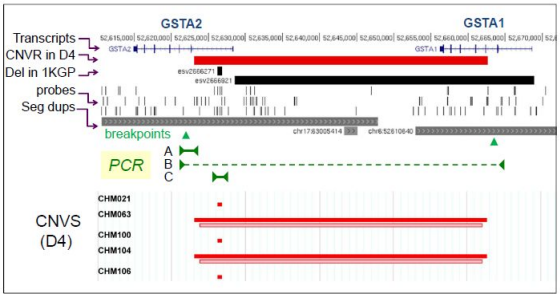

図３　GSTA1-GSTA2 間の NAHR

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．Tahira, T., Yahara, K., Kukita, Y., Higasa, K., Kato, K., Wake, N., Hayashi, K. A definitive haplotype map of structural variations determined by microarray analysis of duplicated haploid genomes. *Genomics Data* （査読有）2, 55-59 (2014)
DOI: 10.1016/j.gdata.2014.04.006

2. Kukita, Y., Uchida, J., Oba, S., Nishino, K., Kumagai, T., Taniguchi, K., Okuyama, T., Imamura, F., Kato, K. Quantitative identification of mutant alleles derived from lung cancer in plasma cell-free DNA via anomaly detection using deep sequencing data. *PloS one* （査読有）8(11), e81468 (2013)
DOI:0.1371/journal.pone.0081468

3. Collin, R. W., Nikopoulos, K., Dona, M., et al. (他 23 名，11 番目) ZNF408 is mutated in familial exudative vitreoretinopathy and is crucial for the development of zebrafish retinal vasculature. *Proc. Natl. Acad. Sci. USA* （査読有） 110(24), 9856-9861 (2013)
doi: 10.1073/pnas.1220864110

4. Kondo, H., Kusaka, S., Yoshinaga, A., Uchio, E., Tawara, A., Tahira, T. Genetic variants of FZD4 and LRP5 genes in patients with advanced retinopathy of prematurity. *Molecular vision*（査読有）19, 476 (2013)
http://www.molvis.org/molvis/v19/476/

〔学会発表〕（計　３件）

1．田平知子，久木田洋児，矢原耕史，山本健，加藤聖子，和氣徳夫，林健志　日本人ハプロイド試料のゲノムワイド解析により同定された構造多型の機能予測．第 35 回日本分子生物学会年会，　2013 年 12 月 5 日，神戸

2．田平知子，久木田洋児，加藤聖子，和氣徳夫，林健志 Structural variations of pharmacogenetic genes detected in haploid genomes of Japanese population．第 72 回日本癌学会学術総会，2013 年 10 月 5 日，横浜

3．田平知子，久木田洋児，矢原耕史，山本健，加藤聖子，和気徳夫，林健志　日本人集団の確定的ハプロタイプ決定とゲノム薬理学への応用．日本薬学会第 133 年会，2013 年 3 月 28 日，横浜

〔その他〕
ホームページ等
D-Haplo データベース
http://orca.gen.kyushu-u.ac.jp/

６．研究組織
(1)研究代表者
田平　知子（TAHIRA TOMOKO）
九州大学・生体防御医学研究所・講師
研究者番号：５０１５５２３０

(2)研究分担者
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）
研究者番号：６０３７２７４４

(3)連携研究者
山本　健（YAMAMOTO KEN）
九州大学・生体防御医学研究所・准教授
研究者番号：６０２７４５２８

(4) 研究協力者
和氣 徳夫（WAKE NORIO）
九州大学・環境発達医学研究センター・特任教授
研究者番号：６００９１５８４

(5)研究協力者
林　健志（HAYASHI KENSHI）
九州大学・生体防御医学研究所・名誉教授
研究者番号：００００１９６７１