様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 １３ 日現在

機関番号：３２６６１
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５４３０１７２

研究課題名（和文）次世代シーケンシングによるヌタウナギの生殖細胞と体細胞の比較ゲノミクス

研究課題名（英文）Comparative genomics of germline and somatic genomes in Eptatretus burgeri through NGS

研究代表者
久保田　宗一郎（KUBOTA, Souichirou）

東邦大学・理学部・教授

研究者番号：３０２７７３４７

交付決定額（研究期間全体）：（直接経費）　　4,000,000円

研究成果の概要（和文）：　円口類のヌタウナギは、染色体放出を行うことが知られている。本研究課題では、次世代シーケンサーを用いてこの種の生殖細胞ゲノムと体細胞ゲノムの全ゲノム配列それぞれを決定した。アセンブルの結果得られたドラフトゲノムは、体細胞で、Scaffold(>500bp) が94,650本、生殖細胞で83,538 本、k-mer (k=40) 頻度分布で補正したゲノムサイズは体細胞で 1.77 Gb、生殖細胞では2.09 Gbと、これまでの解析から想定した値に近似した。現在、RNA-seqデータのマッピング等により、生殖細胞特異的な領域の詳細を解析中である。

研究成果の概要（英文）：In inshore hagfish, Eptatretus burgeri, it is known that chromosome elimination occurs during early embryogenesis, and shown that the eliminated chromosomes are mosaics of highly repetitive, germline-restricted DNA families. In this study, using the next-generation sequencing technology, whole DNA sequences of somatic and germline genomes in this specie were determined. The current assembly constructed 94,650 scaffolds (>500bp) in somatic, and 83,538 scaffolds (>500bp) in germline genomes, subsequently the genome size of somatic and germline genomes were estimated to be 1.77 Gb and 2.09 Gb, from the k-mer (k=40) analysis, indicating the similar values calculated by previous studies. At present, the germline-restricted sequences, namely the eliminated genome is determined and investigated by the mapping of RNA-seq data and annotated the transcriptional regions.

研究分野：分子遺伝学

キーワード：全ゲノム　比較ゲノム　生殖細胞ゲノム　体細胞ゲノム　染色体放出　ゲノム再編成　次世代シーケンシング　ヌタウナギ

１．研究開始当初の背景

一般に高等生物は、体を構成するどの細胞も等しい染色体数（DNA 量）を持つが、幾つかの生物群では体細胞と生殖細胞の間で染色体数や DNA 量が大きく異なることが知られている。これは、発生初期に 2 つの細胞系列が分化する過程において、始原体細胞になる割球から染色体が失われることによる。この現象は染色体放出（染色体削減）と呼ばれて動物界に広く観察されており、後口動物では脊椎動物円口類ヌタウナギ目ヌタウナギにおいて初めて観察された（Kohno *et al.*, 1986）。以来我々は観察したヌタウナギ目魚類 8 種すべてにおいて染色体放出を確認し、染色体全体が放出される場合と染色体末端部分だけが放出される場合があること、放出染色体は主にヘテロクロマチンであること、更に生殖細胞に特異的に増幅した高頻度反復配列を 16 種類同定した（Kubota *et al.*, 1993;他）。

２．研究の目的

これまでの研究から、染色体放出の生物学的役割として１つの可能性が考えられている。それは、生殖細胞と体細胞で働く遺伝子を切り替えるスイッチの役割である。そこで我々は、まず次世代シーケンサーを用いて生殖細胞ゲノム（G ゲノム）と体細胞ゲノム（S ゲノム）をそれぞれ決定し、次に両者を比較解析することで、放出される生殖細胞に特異的なゲノム（放出されるゲノム;E ゲノム）を明らかにし、これまで未知であった染色体放出現象の抜本的な理解と、染色体放出と 2 細胞系列の「分化」との関連性を追求する。

３．研究の方法

(1) G・S 両ゲノムともにペアエンド・データ(G ゲノム：total 321.2Gb, S ゲノム：total 286.9Gb)にメイトペア・データ(G ゲノム：3,5,8,10,12,15,18,20kb の配列, S ゲノム:3,5,8,10kb の配列)を加えてアセンブルを実施し、全ゲノム配列決定を行った。また、リード中の k-mer 頻度の分布を解析し、アセンブル結果におけるリピート配列の割合を調べた。

(2) E ゲノム領域について解析するにあたり、G・S 両ゲノム全体を比較することは作業上負荷がかかり過ぎて困難である。そのため、今回はマッピング解析をベースとした手順で進めた。S ゲノム及び G ゲノムのリード配列を生殖細胞のドラフトゲノムにマッピングし、同一領域中にマップされる G ゲノムリード配列数に対する S ゲノムのリード配列数の割合を比較した。この値が閾値以下の領域を欠失領域つまり E ゲノムとして判定した。尚、閾値については 0%,10%,25%,50%でそれぞれ試行し、マップされた数は正規化やノイズ除去の目的で信頼性の低いもの(40bp より短い配列)は除いて算出した。また、S ドラフトゲノムについても同じマッピング作業を行い、G ゲノムで欠失している領域が存在するかどうかを確認した。検出された欠失領域の詳細は、NCBI のデータベースに BLASTX で検索して調べた(E-値<1e-5)。

(3) 同時に k-mer 出現頻度解析をベースとした欠失領域の検出も行った。手順として、G・S 両ゲノムのリードから k=40 の出現回数を求め、推定コピー数の差が 100 回以上(欠失領域と判定)の 40-mer を抽出した。このうち 39 塩基が共通しているものを繋いでいき、得られた配列中から重複部分を除去してタンデムリピートを検出する。検出されたものの中で、互いに相同性をもつ配列は CD-hit を用いてクラスタリングを行った。推定コピー数は、各 40-mer の出現回数をピークに対応する値で割ることで、ゲノム中での 40-mer のものを算出した。

４．研究成果

(1) k-mer 出現頻度(k=32)は、全リード中に G ゲノムで 103 回、S ゲノムでは 108 回出現していたので、今回の NGS データでは G ゲノムを平均 103 倍、S ゲノムでは 108 倍の厚みで読めていることが明らかとなった。

アセンブルに関して、G ゲノムは Scaffold (>500bp) 83,598 本、全長 1.6Gb という結果が得られた。これに対して、S ゲノムでは Scaffold(>500bp) 94,650 本、全長 1.7Gb となった。このドラフトゲノム構築の結果は、体細胞の方が生殖細胞よりもやや大きいという、想定とは大小関係が逆転した。また、G ゲノムは S ゲノムよりも入力シーケンスライブラリーが豊富であるのにも関わらず、N50 長が 1Mb に達しなかった。K-mer 頻度分布解析(k=32)の結果、G ゲノムにおけるリピート配列率は S ゲノムより大きく 0.426 と既報の他種ゲノム(シーラカンスでは 0.252、Nikaido *et al.*,2013;他)を凌駕しており、これがアセンブルの障害となったと考えられた。そこで、k-mer(k=40)頻度分布からドラフトゲノムに N が含まれる割合を推定し、それをもとにゲノムサイズを再度推定した結果、体細胞ゲノムは 1.77Gb、生殖細胞ゲノムは 2.09Gb となり、想定ゲノムサイズ（1.7、2.17Gb）に近似した。

(2) ① マッピング解析をベースとした手順において検出された欠失領域の推定サイズを次の表に示す。

閾値	0%	10%	25%	50%
S ゲノムで欠失	3.6 Mb	5.4 Mb	9.3 Mb	33 Mb
G ゲノムで欠失	2.1 Mb	4.0 Mb	6.1 Mb	20 Mb

我々が検証すべき E ゲノムは、生殖細胞が体細胞に分化する過程において放出される

ものであり、S ゲノムでのみ欠失領域が認められると想定したが、G ゲノムにおいてもこれが検出された結果となった。本領域の存在を証明すべく、40bp 未満と決めた領域長の下限をいくつか定め直した上で G・S 両ゲノムにおいて欠失と判定される領域長を確認した。閾値 0%でのそれらを次の表に示す。

領域長の下限 (bp)	S ゲノムで欠失とされる領域長 (bp)	G ゲノムで欠失とされる領域長 (bp)
40	3.6Mb	2.1Mb
100	2.5Mb	1.5Mb
150	2.3Mb	1.3Mb
1000	0.55Mb	0.25Mb

G ゲノムにおける欠失領域を仮にノイズであるとした場合、S ゲノムでの欠失の半分以上もノイズであると推測されてしまい、これでは解析の精度が危ぶまれてしまう。さらに、領域長の下限を 1000bp まで上げても依然検出されたことから、単なるノイズではないであろうと考えられる。そこで、G ゲノムで欠失と判定される領域についても個別に観察した。

② G ゲノム中で欠失と判定された領域を長いものから観察していったところ、G ゲノムの欠失領域は S ゲノムではヘテロである例が複数確認された。これらの BLASTX による検索結果には、ウイルス由来等の外来配列である可能性があるものが存在した。これについては、体細胞系列で挿入された配列は生殖細胞系列には存在せず、相同染色体組の一方のみに挿入されるためであると仮説する。本来の研究目的であった、S ゲノム中で欠失と判定された領域における BLASTX 検索結果の概略を次に示す。

配列長(3.6Mb)　閾値 = 0%
　ヒットが存在：3%
トップヒットの生物種
顎口上綱：57%　前口動物：27%　その他：16%

対照として、ランダム抽出した非欠失配列についても BLASTX 検索した。

配列長(3.6Mb)
　ヒットが存在：10%
トップヒットの生物種
顎口上綱：56%　前口動物：30%　その他：14%

欠失配列と非欠失配列でトップヒットする生物種の分類には大きな違いは認められなかった。アミノ酸データベースへのヒットが存在することから、E ゲノムにはタンパク質コード領域も含まれる可能性が示唆された。

(3) k-mer ベースの欠失領域検出のうち、39 塩基が共通しているものを繋いで得た結果は次の通りである。

配列種類数	最大長	推定欠失塩基数
791 本	360bp	320Mb

推定欠失塩基数は、出現回数分布から推定した欠失ゲノムサイズである 300Mb に近い値となった。これらのクラスタリング等の結果は以下の通りである。

791 本→130 種類の配列
各クラスタを反復ファミリーとした
各ファミリーの欠失コピー数
1.4×10^{2}～2.1×10^{6}

出現回数の差が多い 40-mer のトップ 100 は、全て今回分類された 130 種のファミリーに属することも確認した。本ファミリーには我々の先行研究において確認されたものと同じ配列も認められた。その対応表を以下に示す。

		推定欠失コピー数		
配列名	長さ	先行研究	本研究	関連性
EEEb1	64bp	5.5×10^{6}	2.1×10^{6}	一致
EEEb6	56bp	1.0×10^{6}	8.8×10^{4}	1bp 違い
EEEb4	67bp	9.7×10^{5}	5.1×10^{4}	一致
EEEb2	57bp	3.9×10^{5}	4.8×10^{3}	一致
EEEb5	58bp	8.9×10^{4}	4.3×10^{3}	1bp 違い

本研究で推定欠失コピー数が最大であったファミリーは、先行研究において主に欠失したとされる EEEb1 と完全に一致した。また、先行研究では未報告であったファミリーも今回検出された。

(4) 今回の結果に対する課題及び展望について

① アセンブルによって得られたドラフトゲノムは公表するに値する水準であるが、今後の更新が必要かを検討する。特に RNA-Seq データのマッピングによる転写領域のアノテーション作業は必要である。

② マッピングベースによる欠失領域の検出において、ウイルス様外来配列候補が示されたが、全体に占める割合や詳細な構造は不明であり、これを証明するには更なる解析が必要である。

③ k-mer ベースによる欠失領域の検出については、コピー数の差が 100 に満たない配列やリピートファミリ内で代表配列と相同性の低い配列等、今回の方法では検出できない欠失領域も存在する。これは今回の推定欠失量と先行研究の DNA 欠失量との差の原因になっている可能性がある。検出方法に改善の余地があるかの検討が必要である。

５．主な発表論文等

なし

６．研究組織

(1)研究代表者

久保田　宗一郎（KUBOTA, Souichirou）
東邦大学・理学部・教授
研究者番号：３０２７７３４７

(2)研究分担者

後藤　友二（GOTO, Yuji）
東邦大学・理学部・講師
研究者番号：７０３６２５２２

(3)研究協力者

桑田　祐輔（KUWATA, Yusuke）
東邦大学・理学部・博士研究員
研究者番号：００７７２７０８