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 ２７ 年　 ６ 月 １７ 日現在

機関番号：６３８０１
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００３６６

研究課題名（和文）新型シーケンサ・アーカイブ配列を用いた植物系統ＳＮＰ統合と多様性指標解析

研究課題名（英文）Integration of plant strain SNPs using NGS Sequence Read Archive and diversity statistics analysis

研究代表者
神沼　英里（Kaminuma, Eli）

国立遺伝学研究所・生命情報研究センター・助教

研究者番号：９０３１４５５９

交付決定額（研究期間全体）：（直接経費）　　4,100,000円

研究成果の概要（和文）：本研究では、新型シークエンサのアーカイブ配列データベースSequence Read Archiveを用いて、植物系統の一塩基多型を統一基準で統合解析し、DNAPod(http://tga.nig.ac.jp/dnapod/)に統合SNPデータベースを、DDBJ Pipeline(http://p.ddbj.nig.ac.jp/)の一機能として解析ワークフローを構築した。統合SNPデータは、形質マッピング等のゲノム育種研究や集団多様性研究の基盤資源になる。形質マッピングの遺伝マーカとして統合SNPデータを利用する事が可能になり、集団毎に多様性指標解析も行うことが出来る。

研究成果の概要（英文）：In this study, an integrated database of Single Nucleotide Polymorphisms of plant strains from Sequence Read Archive was constructed as DNA Polymorphism Annotation Database(DNAPod). The integrated database is accessible from DNAPod website(http://tga.nig.ac.jp/dnapod/). Moreover, an analytical workflow for DNAPod database was implemented as a function of DDBJ Pipeline (http://p.ddbj.nig.ac.jp/). The constructed DNAPod database can be a data resource of further genomic breeding research and genomic diversity research. The integrated SNPs of rice, sorghum and maize in DNAPod are available for experimental researchers as genetic markers such as trait mapping and diversity statistics analysis.

研究分野：バイオインフォマティクス

キーワード：SNP　次世代シークエンサ　Sequence Read Archive　多様性指標　データベース　解析ワークフロー　一塩基多型　WGS

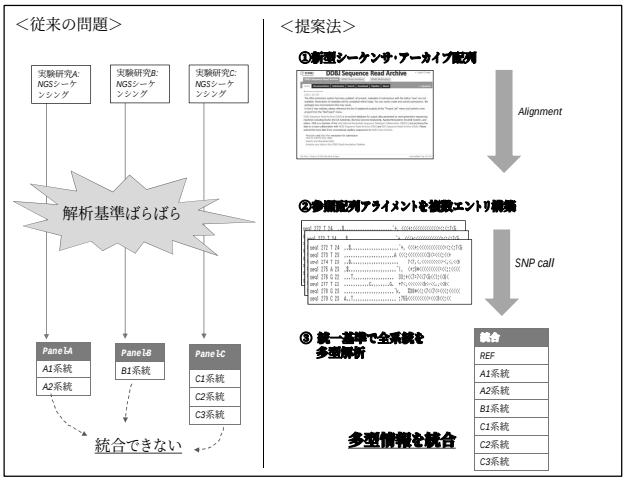
１．研究開始当初の背景

植物の遺伝的多様性研究では1001 genomes (Cao et al., NatGen43:956,2011)等の、新型シークエンサ(New Generation Sequencer : NGS) 利用の自然集団の全ゲノム解読プロジェクトが進んでいる。特にゲノムワイドに形質関連マーカを探索する用途で、SNP genotyping arrayを構築する研究が相次いで発表されている。多型は対象集団により異なるので、SNPパネルの内容は選択した系統に依存する。本研究ではNGSを用いた高密度SNP解析(従来)に、データベース統合の網羅性を付加（新規）する事で、高精度な植物系統高密度SNPデータベースの構築を目指す。

研究代表者は日本DNAデータバンク(DDBJ)がサポートするNGS配列アーカイブのDRA解析パイプライン(DDBJ Pipeline)を構築している(Kaminuma et al.,NAR 38:D33,2010)。また申請者は形質マッピングプログラムの開発経験がある為に、形質責任遺伝子の特定には高精度な系統コレクション用SNPパネルの構築が重要と考えている。パイプラインは大量配列をハイスループット処理する機能を持つ。高精度な統合SNPパネルの構築は、分子遺伝研究者に有益と考えた。

２．研究の目的

植物分野では、次世代シークエンサ由来の系統コレクション高密度SNPパネルが続々と発表されており、タグSNP解析等で絞り込まれた後で、遺伝子型タイピング実験用にSNPアレイにまとめられたりしている。これら高密度SNPパネルは実験研究単位の系統でまとまっており、追加系統分の新規SNPは統合するしかないが、解析基準がばらばらで統合処理は一般に困難である。そこで解析対象となる全系統の新型シークエンサのアーカイブ配列を使って、高密度でSNPを再解析し直して、全系統コレクションの統合SNPパネルを構築する方法を提案する。本研究の目的は、新型シークエンサのアーカイブ配列データベース(DDBJ Sequence Read Archive:DRA)を用いて、統一基準で複数パネル分のSNPの再解析を行い、ゲノム多様性研究の知識基盤となる統合SNPパネルを提供する事である。



３．研究の方法

初年度は「SNPの統合解析の確立」を目指す。DRA公開済の植物登録系統でin silicoで統合SNPパネル(putative SNPs)を構築する。構築方法は、まずDRAのゲノム研究データから、系統単位でFASTQ形式の生配列を抽出する。次にゲノムへのアライメントを行い標準解析ツールでSNP検出を行う。この時、塩基毎のQualiyScoreやアライメント深度を記載した中間ファイルを保存しておき、検出精度の変更に対応可能にしておく。系統毎のSNPsの和集合を取り、統合SNPsとする。

2年度目は統合SNPについて、SNP毎にアリル頻度(Minor Allele Frequency:MAF)と連鎖不平衡(Linkage Disequilibrium: LD)を多様性指標として求める。集団多様性解析や有用形質の責任遺伝子探索に、MAFとLDは重要な統計量である。

3年度目は、「SNP注釈ワークフロー構築」を行う。遺伝子領域注釈付けは、有用形質との関連がついている場合に遺伝子領域構造と共に、系統間差を確認できる機能である。遺伝子領域構造のSNP注釈ワークフローは、DDBJ pipelineへ実装する。DDBJ pipelineは参照配列に

マッピングできる基礎処理部と、マッピング後の高次処理ワークフローに分かれている。高次処理部は米Penn State UnivのGalaxy interface (Giardine et al.,Genome Res, 2005)でワークフローが構築できる。これまでも多型注釈やRNA-seq解析のワークフローを構築してきた。ここに統合SNP用のワークフローを追加する。

4. 研究成果

□ **研究成果①: DRA公開済の植物系統でSNP統合解析を確立**

初年平成24年度の成果として、DRA公開済の植物登録系統で統合SNPパネルを構築した。ゲノムが公開されておりDRA登録数が多い植物種としてイネを選択した。Whole Genome Shotgun研究のDRA植物エントリから678のイネ系統を抽出して、SNP解析を行い統合した。データ構築方法は、DRAのゲノム研究データから系統単位でFASTQ形式の生配列を抽出して、参照ゲノム配列Oryza sativa 日本晴系統(IRGSP build05)へのアライメントを行い、samtools解析ツール(Li et al., Bioinformatics,15:2078-9,2009)でSNP検出を行った。SNP数が多かったOryza nivara IRGC 105327系統(SRS086324, SNP数:N=3.5x10^6)から、最小数の日本晴系統(ERS006293,N=1.5x10^3)まで同一解析基準でSNPを構築した。678系統のうち、621系統は深度5x以下の薄読データである。

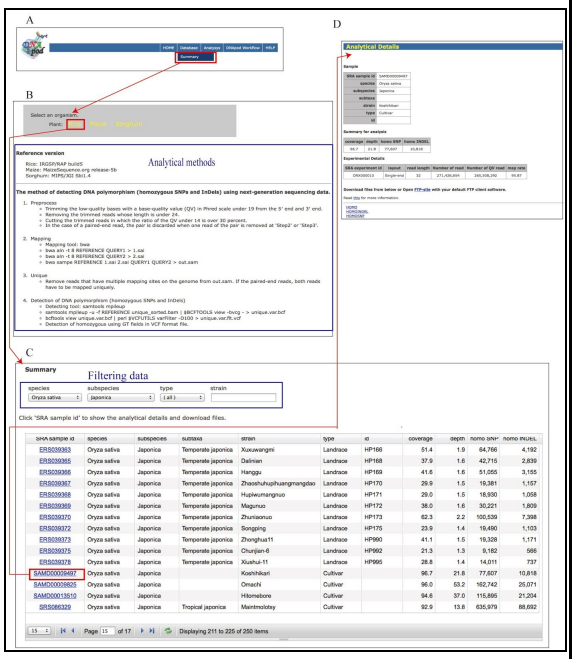
□ **研究成果②: 統合SNPの精度評価と多様性指標解析**

統合用の解析パラメータの検討として、SNP計算の為のゲノム被覆率と深度の関係を調べた。DRA公開データとして、DRA000307のイネ系統の全ゲノムシークエンシングデータを用いた。日本晴ゲノムIRGSP build05参照配列に対して、DRA000307配列をアライメントして、シークエンシング・データ量に依存する深度と参照配列被覆率を計算した。またRice Annotation Project(RAP)の34,792遺伝子について遺伝子数被覆率も計算した。結果として、参照配列被覆率と遺伝子数被覆率も、共に深度15～20程でほぼ一定状態になった。これにより配列アーカイブの解析条件として、深度15以上を一つの目安とする事が出来る。一方、多様性統計指標解析については、テスト的に柑橘11品種全ゲノム配列実験データについて、アリル頻度解析と連鎖不平衡解析を行った。事例結果として、bi-allelic SNPsのみ抽出した柑橘データでMinor Allele Frequency(MAF)を計算して、MAF>0.05で残ったSNPの割合は約75%だった。最終的に、ヘテロ接合度期待値(Expected Heterozygosity:He)、アリル頻度(Allele Frequency), 多型情報量(Polymorphism Inforamtion Content: PIC)の計算プログラムを構築した。

□ **研究成果③: DNApodデータベースと統合SNP注釈ワークフローの構築**

統合SNP注釈ワークフローを、NGS配列注釈ツール「DDBJ Read Annotation Pipeline (http://p.ddbj.nig.ac.jp/)」の高次処理部に1機能として実装した。ワークフローではSnpEffツール(Cingolani P,Fly, 6:80, 2012)により、遺伝子の構造領域情報を基に、対象SNPの効果注釈（非同義置換等）を行う。一方、イネ678系統の多型データは初年度にDNA Polymorphism annOtation Database (DNApod) (http://tga.nig.ac.jp/dnapod/)として公開した。最終年度には、DRA登録WGS由来のイネ679系統、ソルガム66系統、トウモロコシ404系統の3植物属全1,149系統の多型データをDNApodに登録した。1,149系統の多型データのうちホモ接合SNPについて効果注釈を行い、注釈結果をDNApodサイトからダウンロード可能にした。下図にその結果を示す。DNApodウェブサイトのメニュ(図A)からDatabase→Summaryを選択すると、生物種の選択画面(図B)になる。選択すると、DRAに

登録されている系統単位のサンプル番号が出る(図 C)ので、選択すると解析結果統計とダウンロードのリンクが現れる(図 D)。



以上のように本研究では、新型シークエンサのアーカイブ配列データベース DRA を用いて、植物系統の一塩基多型を統一基準で統合解析し、ゲノム多様性研究の知識基盤として DNAPod データベースと DDBJ Pipeline に解析ワークフローを構築した。統合 SNP データは、形質マッピング等の遺伝育種研究や集団多様性研究の基盤資源になる。形質マッピングの遺伝マーカとして統合 SNP データを利用する事が可能になり、集団毎に多様性指標解析も行うことが出来る。

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 1 件）

DNApod: DNA polymorphism annotation database from next-generation sequence read archives", Takako Mochizuki, Yasuhiro Tanizawa, Takatomo Fujisawa, Naruo Nikoh, Tokurou Shimizu, Atsushi Toyoda, Asao Fujiyama, Nori Kurata, Hideki Nagasaki, Eli Kaminuma, Yasukazu Nakamura, BMC genomics, submitted.

〔学会発表〕（計 9 件）

1. Mochizuki T, Tanizawa Y, Fujisawa T, Nikoh N, Toyoda A, Fujiyama A, Kurata N, Nagasaki H, Shimizu T, Kaminuma E, Nakamura Y, DNA Polymorphism Database from New-Generation Sequence Read Archive and Analytical Workflow,Plant & Animal Genome XXIII , San Diego, 2015.1,
2. 望月孝子，谷澤靖洋，藤澤貴智，長崎英樹，神沼英里，清水徳朗，豊田敦，藤山秋佐夫，倉田のり，二河成男，中村保一，DNApod：NGS アーカイブ配列からの統合 DNA 多型注釈データベース 生物横断的解析への活用，第 37 回日本分子生物学会年会，横浜,2014.11.
3. 神沼英里、望月孝子、長崎英樹、谷沢靖洋、小笠原理、大久保公策、高木利久、中村保一，遺伝研スパコン概要と、NGS アーカイブデータを用いた SNP 解析 ,日本育種学会第 126 回講演会「NGS 使い倒し講座 - Breeding Informatics 研究 XIII」,宮崎,2014.9.
4. Takako Mochizuki, Takatomo Fujisawa, Yasuhiro Tanizawa, Hideki Nagasaki, Eli Kaminuma, Tokurou Shimizu, Atsushi Toyoda, Asao Fujiyama, Nori Kurata, Naruo Nikoh and Yasukazu Nakamura, Constructing an integrated DNA polymorphism database and an analytical workflow for plants and bacteria, International Plant & Animal Genome XXII , San Diego, 2014.1.
5. 望月 孝子，藤澤 貴智，谷澤 靖洋，長崎 英樹,神沼 英里，大柳 一,清水 徳朗，豊田 敦，藤山 秋佐夫,倉田 のり,二河 成男，中村 保一，DNA 多型統合データベースと解析ワークフローの構築 ： 植物、微生物への取り組み，第 36 回日本分子生物学会年会，神戸，2013.12.
6. 神沼 英里，藤澤 貴智，望月 孝子，谷沢 靖洋，豊田 敦，藤山 秋佐夫，倉田 のり，清水 徳朗，中村 保一，NGS 由来ゲノムワイド多型マーカ構築とその RDF 注釈情報統合化，

第 36 回日本分子生物学会年会,神戸, 2013.12.

7. 望月 孝子,長崎 英樹,神沼 英里, 大柳一, 清水 徳朗, 豊田 敦, 藤山 秋佐夫, 倉田のり, 二河 成男, 中村 保一, 新型シーケンサアーカイブ配列からの DNA 多型統合データベース DNA Polymorphism annOtation Database (DNApod) の構築, 第 35 回日本分子生物学会年会, 福岡, 2012,12
8. 神沼英里 , 望月孝子, 長崎英樹, 児玉悠一, 猿橋智, 大久保公策, 高木利久,大柳 一,倉田のり, 清水徳朗, 中村 保一, 新型シークエンサのアーカイブ配列と解析パイプライン:系統間 SNP 解析を事例として,第 122 回日本育種学会講演会ワークショップ 06 "育種のための情報解析ツール使い倒し塾"(招待講演), 京都, 2012.9
9. 望月孝子, 長崎 英樹, 神沼 英里, 大柳 一, 清水徳朗, 豊田 敦, 藤山 秋佐夫, 倉田 のり, 二河 成男, 中村 保一, 新型シーケンサアーカイブ配列からの DNA 多型統合データベースと解析ワークフローの構築, 第 123 回日本育種学会講演会,東京, 2012.3.

〔図書〕(計１件)

1.実験医学別冊「次世代シークエンス解析スタンダード」長崎英樹,望月孝子,谷沢靖洋,神沼英里, 中村保一 (担当:共著, 範囲:VI-4 DDBJ Read Annotation Pipeline 解析パイプラインによる RNA-Seq de novo assembly とイネ多型解析) 羊土社, pp.352-360, 2014.8 ,

〔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：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://tga.nig.ac.jp/dnapod/
DNA Polymorphism Annotation Database(DNAPod)

６．研究組織
(1)研究代表者
神沼英里 (KAMINUMA, Eli)
国立遺伝学研究所・生命情報研究センター・助教
研究者番号： 90314559

(2)研究分担者

(3)連携研究者
長崎英樹 (NAGASAKI, Hideki)
国立遺伝学研究所・生命情報研究センター・特任研究員
研究者番号： 70624451

望月孝子 (MOCHIZUKI, Takako)
情報・システム研究機構・新領域融合研究センター・特任研究員
研究者番号： 40709284