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：２４５１０２７３

研究課題名（和文）クラウド型高速シーケンシングアノテーションシステムの開発研究

研究課題名（英文）The Development of a cloud computing-based annotation system for next-generation sequencing data

研究代表者
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

交付決定額（研究期間全体）：（直接経費）　　4,300,000円

研究成果の概要（和文）：近年登場した次世代シーケンサーと呼ばれるDNA解読装置の登場で、ゲノム科学や分子生物学に必要なゲノム/遺伝子の塩基配列の解読能力は飛躍的に向上した。しかし、その膨大な配列データを解析する計算機システムや情報解析を行う人員の不足という事態がおこった。我々は以上の問題を解決するためにDDBJ Read Annotation Pipeline(通称DDBJパイプライン)を開発、インターネット上で公開している。
DDBJパイプラインは、計算機処理のためのコマンド操作が苦手な生物学の実験研究者でもパソコンのマウス操作で扱えるように設計され、解析処理は国立遺伝学研究所のスパコンシステムを利用している。

研究成果の概要（英文）：High-performance next-generation sequencing (NGS) technologies are advancing genomics and molecular biological research. However, the immense amount of sequence data requires computational skills and suitable hardware resources that are a challenge to molecular biologists. The DNA Data Bank of Japan (DDBJ) of the National Institute of Genetics (NIG) has initiated a cloud computing-based analytical pipeline, the DDBJ Read Annotation Pipeline (DDBJ Pipeline), for a high-throughput annotation of NGS reads. The DDBJ Pipeline offers a user-friendly graphical web interface and processes massive NGS datasets using decentralized processing by NIG supercomputers currently free of charge.

研究分野：ゲノム情報解析

キーワード：ゲノム　次世代シーケンサー　解析パイプライン　クラウド　RNA-Seq　ChIP-Seq

１．研究開始当初の背景

(1) 近年開発された高速DNAシーケンサーは、ゲノム解析や遺伝子発現解析などに新た進歩をもたらすとされている。一回の解析で6千億bpにおよぶ解読量は、時間やコストの短縮につながり、近年では大規模なシーケンスセンターだけではなく、より小規模な研究グループでも利用され始めた。しかし、その膨大な塩基配列のデータ量と、一度に解読される配列長(リード長)が100-400bpという短さでしかないため、参照配列へのマッピングや新たに解読された生物の塩基配列のアセンブリといった一次的な解析において、多大な計算機資源の投入と運用に専門知識を有する人材が必要とされる。

研究代表者らが参画しているDNA Data Bank of Japan (DDBJ)は、米国National Center for Biotechnology Information (NCBI)と欧州European Bioinformatics Institute (EBI)と共同で国際塩基配列データベース(INSDC)を構築し、高速シーケンサー由来の配列の登録業務と公開を行っている。このため、解析するための計算機リソースを提供することで高速シーケンサーの利用者の研究業務が促進し、登録数の増加することを期待して高速シーケンサーの解析パイプライン(DDBJ Read Annotation Pipeline: 以下DDBJパイプライン)を構築し、インターネット上で公開し利用に供している(http://p.ddbj.nig.ac.jp/ ①)。

(2) インターネットを介したクラウドコンピューティングと分散処理の技術、そして現在高速シーケンサーのデータ解析で代表的に使用される 12 のバイオインフォマティクスツールを内包して、DDBJパイプラインはリード配列の参照ゲノム配列へのマッピングや *de novo* アセンブリを行い、上記の問題を解消に努めている。

２．研究の目的

(1) DDBJ パイプラインは、膨大な計算量を処理できる環境を、計算機リソースが整っていない研究グループに提供する。しかしながら研究結果に至る一行程を肩代わりしているのにすぎないため、新たに解析手段を追加する必要があった。例として、RNA-Seq 配列の *de novo* アセンブル後の配列にどのような遺伝子がコードされているか調べる機能である。

(2) 個人の塩基配列など使用が制限されているデータに関して、バーチャル・マシン(VM)と呼ばれる情報処理技術を用いてユーザーがDDBJパイプラインを任意の環境(計算機サーバー)で実行できるようにする。

３．研究の方法

(1) 構築済みの DDBJ パイプラインのシステムを基礎解析部とし、参照ゲノム配列へのマッピングや、*de novo* アセンブリといった処理が一元的でありながら計算機負荷が高い処理を担当させた。その結果を引き継いで解析する行程を高次解析部としてwebアプリケーションGalaxy (https://usegalaxy.org/ ②)を基にして構築し、解析用ツールを作成、追加した。

(2) 構築したシステムを平成 24 年度より運用されている国立遺伝学研究所のスパコンシステム②に移植し、解析処理をスパコン上で並列処理し高速化を行った。

(3) VM 化は BioLinux (http://environmentalomics.org/bio-linux/)というパッケージプログラムにツール等を移植した。

４．研究成果

(1) 高次解析部を追加し、現在の構成は

図 1 の通りである。基礎解析部から高次解析部へデータを移行する機構を作成し、続けて作業が可能である。
(2) 国立遺伝学研究所のスパコンシステム上の並列計算処理によって単一の計算機で運用されていた時より解析時間(特に Galaxy を利用した高次解析部で)が大幅に短縮された。平成 27 年 5 月現在の登録ユーザー数は 821 名となっている。

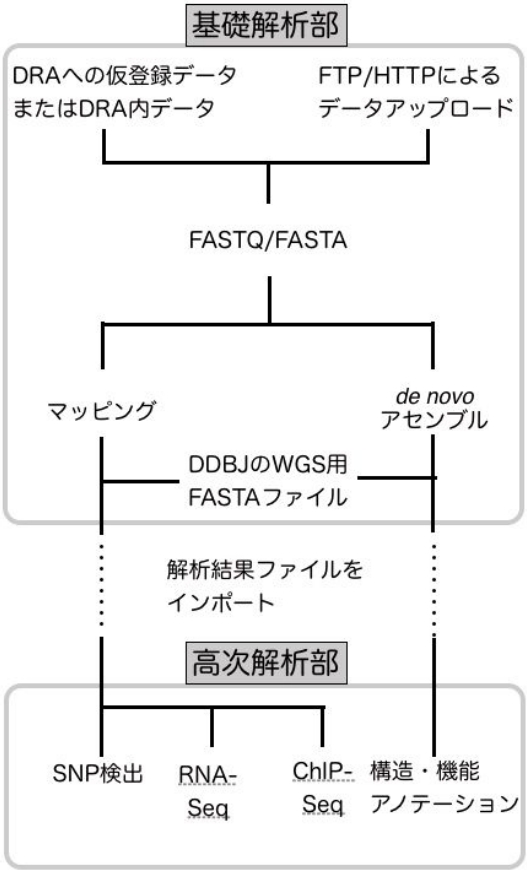

図 1 DDBJ パイプラインの構成

(3) 高次解析部の基礎となっている Galaxy は、個々のユーザーが開発したツールを追加して容易に共有する仕様になっている。基礎解析部の結果を受けて解析するツールとしてゲノムへのマッピング結果から 1)ゲノム上の SNP の分布を図示するツール 2)遺伝子発現量を正規化するツール 3)DNA 結合タンパク質の結合部位候補の同定を行うツール等を追加している。また、RNA-Seq の *de novo* アセンブリ結果からコードされている遺伝子の構造を解析するツール等を追加した。さらに、高次解析単独で動作するツールとしてヒト白血球抗原(HLA)解析ツール④を追加した。
(4) 現在 VM パッケージ化したプログラムファイルを DDBJ パイプラインのホームページより公開している。
(5) 研究成果としては論文、書籍での発表の他、DDBJ 等が主催する講習会で利用法の紹介の講習も行った(学会発表の項参照)。

引用文献

① 長崎 英樹, 神沼 英里: "DDBJの塩基配列解析ツールと登録システム",実験医学増刊 Vol.29 No.15 使えるデータベース・ウェブツール, 2011 2537-43.
② Goecks *et al.* Genome Biol. 2010;11:R86.
③ Ogasawara *et al.* Nucleic Acids Res. 2013;41:D25-9.
④ Hosomichi *et al.* BMC Genomics. 2014 Aug 4;15:645.

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 2 件）

① Kazuyoshi Hosomichi, Shigeki Mitsunaga, Hideki Nagasaki, Ituro Inoue A Bead-based Normalization for Uniform Sequencing depth (BeNUS) protocol for multi-samples sequencing exemplified by HLA-B.
BMC Genomics. 2014, 15:645. 査読有
doi:10.1186/1471-2164-15-645

② Nagasaki H, Mochizuki T, Kodama Y, Saruhashi S, Morizaki S, Sugawara H,

Ohyanagi H, Kurata N, Okubo K, Takagi T, Kaminuma E, Nakamura Y.

DDBJ Read Annotation Pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data
DNA Res. 2013, 20(4):383-390. 査読有
doi: 10.1093/dnares/dst017

〔学会発表〕(計 13 件)
[国内]
① GalaxyWorkshopTokyo2015
東京大学先端科学技術研究センター、東京
4月
長崎英樹 HLA解析ツール

② 第35回日本分子生物学会年会(2012)
福岡国際会議場、福岡 12月
長崎 英樹、藤澤 貴智、望月 孝子、猿橋 智、神沼 英里、石崎 公庸、大和 勝幸、河内 孝之、中村 保一
DDBJパイプラインによるゼニゴケゲノム解析とゲノムアノテーションデータベースの構築
③ 第52回日本植物生理学会年会(2011)
東北大学、 宮城 3月
長崎 英樹、望月 孝子、渡邊 成樹、森崎 彰太、児玉 悠一、猿橋 智、高木 利久、菅原 秀明、大久保 公策、神沼 英里、中村 保一
DDBJ Read Annotation Pipeline : 新型DNAシーケンサ由来配列のクラウド型パイプライン
④ 第34回日本分子生物学会年会(2011)
パシフィコ横浜、神奈川 12月
長崎 英樹、望月 孝子、児玉 悠一、猿橋 智、高木 利久、大久保 公策、神沼 英里、中村 保一
新型シーケンサ・アーカイブ配列のクラウド型解析パイプラインDDBJ Pipeline進捗: de novoアセンブル配列注釈ワークフロー

[海外]
⑤ The International Plant & Animal Genome XXI Conference (2013)
San Diego, USA 1月
Nagasaki H, Fujisawa T, Mochizuki T, Saruhashi S, Kaminuma E, Ishizaki K, Yamato KT, Kohchi T, Nakamura Y.
Liverwort Genome Analysis using DDBJ Pipeline and Construction of the Genome Annotation Database
⑥ The International Plant & Animal Genome XX Conference (2012)
San Diego, USA 1月
Nagasaki H, Mochizuki T, Kaminuma E, Kodama Y, Saruhashi S, Toshihisa T, Okubo K, Nakamura Y.
DDBJ Sequence Read Archive and a cloud-computing based annotation tool for new-generation sequencing data
⑦ The 2nd International Conference on the Progress of "1000 Plant & Animal Reference Genomes Project" (2011)
Shenzhen, China 7月
Nagasaki H, Mochizuki T, Kodama Y, Saruhashi S, Takagi T, Okubo K, Kaminuma E, Nakamura Y.
DDBJ Sequence Read Archive and a cloud-computing based annotation tool for next-generation sequencing data
⑧ Plant and Animal Genome XIX Conference (2011)

San Diego, USA 1月
Nagasaki H, Mochizuki T, Watanabe S, Morizaki S, Kodama Y, Saruhashi S, Takagi T, Okubo K, Kaminuma E, Nakamura Y.
DDBJ Read Annotation Pipeline: A cloud based pipeline for high-throughput analysis of next generation sequencing data

[セミナー等]
⑨ 統合データベース講習会 : AJACS十勝(2014)
長崎 英樹
DDBJ Read Annotation Pipeline の紹介と実習
(RNA-Seq配列のde novoアセンブリを中心に) (招待講演)
⑩ イルミナ株式会社 ウェビナー (2013)
中村 保一、長崎 英樹、谷沢 靖洋
DDBJパイプラインによる RNA-Seq配列のde novoアセンブル (招待講演)
⑪ 第 25-28 回 DDBJing 講習会 (2012-2013)
長崎 英樹
DDBJ Pipeline の紹介
⑫ 第164回農林交流センターワークショップ (2012)
長崎 英樹、 望月 孝子
NIGスパコンを利用したNGSアーカイブ配列再利用とクラウド型解析パイプライン実習 (招待講演)
⑬ 統合データベース講習会 : AJACS名古屋 (2012)
長崎 英樹、 望月 孝子
DDBJパイプラインによる高速シーケンスデータ解析 (招待講演)

〔図書〕(計 3件)
①長崎 英樹: "DDBJ Read Annotation Pipeline", 実験医学増刊 今日から使える! データベース・ウェブツール, 2014 Vol.32-No.20 176-177.
②長崎 英樹: "Galaxy", 実験医学増刊 今日から使える! データベース・ウェブツール, 2014 Vol.32-No.20 178-179.
③長崎 英樹, 望月 孝子, 谷沢 靖洋, 神沼 英里: "DDBJ Read Annotation Pipeline",実験医学別冊 次世代シークエンス解析スタンダード, 2014 352-360.

〔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
DDBJ Read Annotation Pipeline
基礎解析部: http://p.ddbj.nig.ac.jp
高次解析部: http://p-galaxy.nig.ac.jp

６. 研究組織
(1)研究代表者
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

(2)研究分担者
( )

(3)連携研究者
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