

特集記事 ワークショップ報告

遺伝研スパコンとコマンドラインでの NGS データ使い倒し講座

神沼英里¹⁾・望月孝子¹⁾・門田有希²⁾・小林正明^{3,4)}・大柳 一^{3,4)}・矢野健太郎^{3,4)}¹⁾ 国立遺伝学研究所生命情報研究センター, 静岡県三島市, 〒411-8540²⁾ 岡山大学大学院環境生命科学研究科, 岡山県岡山市, 〒700-8530³⁾ 明治大学農学部, 神奈川県川崎市, 〒214-8571⁴⁾ JST・CREST, 埼玉県川口市, 〒332-0012

Introduction of NIG supercomputer with command line-based NGS analysis

Eli Kaminuma¹⁾, Takako Mochizuki¹⁾, Yuki Monden²⁾, Masaaki Kobayashi^{3,4)}, Hajime Ohyanagi^{3,4)} and Kentaro Yano^{3,4)}¹⁾ Center for Information Biology, National Institute of Genetics, Mishima, Shizuoka 411-8540²⁾ Graduate School of Environmental and Life Science, Okayama University, Okayama, Okayama 700-8530³⁾ School of Agriculture, Meiji University, Kawasaki, Kanagawa 214-8571⁴⁾ JST CREST, Kawaguchi, Saitama 332-0012

キーワード

Linux, NGS, スパコン, コマンドライン, 大規模解析

はじめに

近年急速に普及した次世代シーケンス (NGS) 技術を利用することにより, 植物育種学分野においても大規模な遺伝解析が行われるようになった. それに伴い, 大量の配列データを扱うための遺伝解析手法がいくつも開発されている. たとえば, RAD-seq (Baird *et al.* 2008), ddRAD-seq (Peterson *et al.* 2012), MutMap (Abe *et al.* 2012), GWAS (Huang *et al.* 2010) などが挙げられる. これらは, ゲノム網羅的に SNP (single nucleotide polymorphism) および INDEL (insertion および deletion) などの変異情報を出力し, それら変異情報を解析することで, 有用遺伝子の同定および遺伝資源の多様性解析などを加速化する. また, 全ゲノム配列が解読されていない非モデル生物種に関しても大規模な変異解析が可能になりつつある. つまり, NGS の普及はこれまでと比較にならないほど大量の配列データを扱うことを可能にし, これら大規模解析を通じた育種学分野のさらなる発展が見込まれている.

NGS データを扱ったことのない研究者にとって, これら大量の配列データを扱うことは容易ではない. まずは相応の解析環境を準備しなければならない. また, 実際に自身の手で解析を行うことがデータ解析の習得には不可欠である. そこで, 今回は, 実習形式のワークショップ

を企画・開催した. 本ワークショップでは, 解析環境のセットアップに関する解説, さらに, 実際に育種研究でよく用いられている NGS データ解析 (リシーケンスによる SNP 同定およびその可視化, *de novo* のアセンブル解析など) の講習を行った. 参加者は, 必要に応じて自身のパソコンを持ち込み, 試験的にデータ解析を行った. 以下, 3 つの演題に関する概要を解説する. なお, 本文中の網掛けはキーボード入力 (コマンドライン入力), 「」は空白を表す. 長いコマンドラインでは, 本文中で網掛けが複数行に渡るので, 注意してほしい.

演題 1 Linux マシン使用法

矢野健太郎 (明治大・農, JST・CREST)

バイオインフォマティクス解析を実施する上で, Unix (Macintosh) や Linux の利用が極めて有用である. ターミナル (端末) において使用できるコマンドが充実しているため, 特別なプログラムを組まなくても簡便な処理がただちにできる. たとえば, 配列や遺伝子発現などの実験データを格納したテキスト・ファイルから目的の情報のみを抽出するなどが可能である. Linux は, Linux OS をパーソナル・コンピューターなどにインストールすることで直ちに使用できる. スタンドアロンで大きな計算処理を実行しないのであれば, 安価な PC で十分である. 無料で提供されている Linux OS として, CentOS や Fedora などがある. Linux OS のインストール後には, セキュリティの設定を強く推奨する. たとえば, sendmail (メール・サービス) や httpd (Web サービス) など実際には使

用しないプログラムがあるならば、それらは自動起動しないように設定をすべきである。サーバー機能の設定は、「CentOS 6 で作るネットワークサーバ構築ガイド」（サーバ構築研究会、秀和システム）などに詳しく紹介されている（記載した書籍は絶版であり、次のバージョンに対応した新書の発刊が待たれる）。

コマンドラインでの解析は、端末（Macintosh ではターミナル）を通して行う。実際の解析は、スタンドアロン型とサーバー・クライアント型に大別される。スタンドアロン型では、BLAST などのプログラムを自分自身（ローカル）のマシンにインストールし、端末を通してインストール済みのプログラムを実行する。ローカル・マシンがもつ CPU やメモリ、ストレージ（ハードディスクなどの記憶領域）などの計算機リソースを使用するため、処理速度などはローカル・マシンの処理能力に依存する。ネットワークを介さないため、ネットワークがない環境でも実行できると共に、データの機密保持において頑健である。サーバー・クライアント型では、一般的には、解析サーバーにユーザー名とパスワードを使って SSH（または、Telnet）ログインし、解析サーバーにインストールされているプログラムを実行する。実行処理では、ローカル・マシンではなく、解析サーバーの計算機リソースを使用する。そのため、一般ユーザーは、解析処理を行うマシン（解析サーバー）の購入・更新・管理・運営といった作業や設備投資から解放される。その一方で、解析サーバーの管理者ではないために、実行したいプログラムが解析サーバー上に存在しない場合、その都度、インストールを依頼する必要がある（もちろん、管理上の理由などから、インストールしてもらえないこともある）。また、解析に使用するデータ・ファイルがローカル・マシンに存在する場合、データ・ファイルを scp コマンドなどでローカル・マシンから解析サーバーに転送する必要がある。同様に、解析結果のデータ・ファイルについてもローカル・マシンで処理する場合（エクセルで作図するなど）には、解析サーバーからローカル・マシンに該当ファイルを転送する。

端末を用いた解析を実行する上で、初歩的なコマンドや操作法を知っておくと効率的である。初歩的な Linux コマンドには、ls, cd, pwd, ps, more, less, cat, head, tail, grep (egrep), cut, join, sort, comm, diff などがある。コマンドの使用法は、解析書籍やオンラインに記述が多く存在する。多くのコマンドは、オプションを有している。たとえば、ファイルやディレクトリのリストを出力する ls コマンドを、lt オプション（えるていー、数字の 1 ではありません）を付加して実行する場合、ls -lt と入力する。この lt オプションでは、ファイルやディレクトリをタイムスタンプ順（より新しいファイルが先）に並び替えて出力する。また、パイプ(|)によって、複数のコマンドをつなげて実行できる。たとえば、ls -lt | head とすると、ls -lt の結果が head コマン

ドに渡されて、最初の 10 行のみが出力される（ls -lt の出力結果が 10 行に満たない場合、すべての出力結果が出力される）。コマンド操作による標準の出力先はモニターとなっている。コマンドの最後に出力ファイルを指定することで、出力結果をファイルに保存できる。たとえば、ls -lt > test.out とすることにより、ls -lt の結果をモニターではなく、ファイル test.out に保存できる。cat や more, less コマンドによりファイルの内容をモニター出力できる（例、cat test.out）。コマンドを使用していくと、何度も同じコマンドを実行するが多い。その際に、ヒストリー機能を用いると便利である。ヒストリー機能では、キーボードの上下矢印キーを押すことによって、以前入力したコマンドを呼び出すことができる。また、コマンドや対象ファイル名をコピー・ペーストで入力することにより、入力を簡便にできる。端末の設定によって方法は異なるが、多くの場合、コピーしたい文字列をドラッグして反転させるだけで、その文字列をコピーできる。ペーストは、マウスのホイールボタンを押すか、右クリックより実行できる。キーボードに触る回数を減らすことにより、タイプミスも少なくなるため、迅速かつ効率的に処理できる。

演題 2 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

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・生命情報）

国立遺伝学研究所（大学共同利用法人 情報・システム研究機構、以下遺伝研）では、生命科学研究者向けの大規模計算機利用拠点として、スーパーコンピュータサービスを提供している（Ogasawara *et al.* 2013）。遺伝研のスーパーコンピュータ（NIG スパコン）は、5 年毎にシステムハードウェアが最新式に更新されており、2014 年 9 月現在のシステムは、2012 年 3 月の初期導入および 2014 年 4 月の中間増強が行われた 3 年目の計算機群である。初期導入分は Phase 1、中間増強分は Phase 2 と区分されており、ゲノム解析用に整備された大規模メモリ 10TB/2TB の計算機（fat/medium ノード）に加えて、数百の 64 GB メモリの計算ノード（thin ノード）を提供している。表 1 に一般利用の研究者向けに提供されている計算機数と性能をまとめる。2012 年 6 月時点で公開された、世界のスーパーコンピュータ性能 Top500 ランキング（<http://www.top500.org/>）では、CPU の理論ピーク性

表 1. NIG スパコンの計算機性能（一般利用向け）

導入時期区分	Phase 1			Phase 2	
	Fat	Medium	Thin	Medium	Thin
ノード数	1	2	199	8	193
CPU コア数/ノード	768	80	16	80	20
メモリ容量/ノード	10TB	2TB	64GB	2TB	64GB

能値 128.4TFlops で 280 位であった。

育種研究者は、NIG スパコンの計算機と共に、内部に設置されているデータベースや解析プログラムツールを無償で利用できる。利用には、まず NIG スパコンのウェブサイト (<http://sc.ddbj.nig.ac.jp/>) から利用申請を行う。海外の居住者は利用制限があるので注意が必要である。申請が許可されれば、郵送でアカウントとパスワードが送られてくる。日本の他のスーパーコンピュータは利用料金がかかるが、NIG スパコンは現時点で無償である（ただし課金の議論はされている）。ターミナルを起動し、次の鍵括弧内のコマンドを入力して、3 ステップでスパコンを使ってみよう。

1. スパコンへログイン「`ssh_kaminuma@gw.ddbj.nig.ac.jp`」: 自分のパソコンから、インターネット越しに NIG スパコンに `ssh` (secure shell) コマンドで、リモートアクセスを行う。アカウント ID (ユーザー名, kaminuma) は自分の ID に変更する。gw.ddbj.nig.ac.jp は Phase 1 用で、Phase 2 の場合は、gw2.ddbj.nig.ac.jp となる。パスワードを聞かれるので入力する (※2015 年 3 月末に NIG スパコンへのログインは、公開鍵認証方式へ移行しました)。
2. ジョブ投入用ノードにログイン「`qlogin`」: gw.ddbj.nig.ac.jp は、ゲートウェイと呼ばれるスパコンの入口計算機に入るの、そこから `qlogin` コマンドでジョブ投入用ノードに入る。
3. ジョブの投入「`qsub -l_short_tmp.sh`」: シェルスクリプト `tmp.sh` を 2 時間終了キュー (`-l short`) で実行する。`tmp.sh` は「`echo `ls -d /usr/local/pkg/*/*_tmp.txt`>tmp.sh`」であらかじめ作っておく。bwa や Cufflinks など解析プログラムのディレクトリ・リストが `tmp.txt` に保存される。終了する際は、「`exit`」コマンドを実行し、ジョブ投入用ノードからログアウトし、再び「`exit`」コマンドを実行することで、NIG スパコンからログアウトする。

ジョブ投入の詳細は、NIG スパコンウェブサイトの基本的な利用方法 (<http://sc.ddbj.nig.ac.jp/index.php/ja-howtouse>) を参照されたい。利用可能なデータベースや解析プログラムの設置場所も、ウェブサイトで確認できる。

また、筆者らは、NIG スパコンを利用した次世代シーケンサの配列解析パイプラインサービス“DDBJ Read Annotation Pipeline (DDBJ Pipeline)”を 2009 年から提供している (Kaminuma *et al.* 2010, Nagasaki *et al.* 2013)。計算機資源不足や解析支援員不足の状況にある実験研究者の支援目的で構築したサービスで、NIG スパコンの遠隔利用と大規模配列解析のワークフロー定型化が特徴である。NIG スパコンでの解析は、プログラムを書く必要がある、コンピュータの扱いに慣れていない実験研究

者には敷居が高い。DDBJ Pipeline では、プログラミングをせずともグラフィカルユーザーインターフェースを利用して次世代シーケンサ配列の解析を実行できる。また、NGS 配列アーカイブ「DDBJ Sequence Read Archive (DRA, Kodama *et al.* 2012)」の公開データを、アクセス番号の入力だけで DDBJ Pipeline に取込む機能がある。さらに SNP 解析ワークフローを使って、DRA 公開データを網羅的に解析した、系統間多型データベースを公開している (<http://tga.nig.ac.jp/dnapod/>)。プログラミング初心者が NIG スパコンで次世代シーケンサの配列解析を進める時に、DDBJ Pipeline で解析結果を出しておく和良好的かもしれない。Pipeline のログに実行コマンドがすべて記載されているので、NIG スパコンで同一のジョブを実行してみれば、解析の練習になるだろう。

演題 3 コマンドラインを用いた NGS 解析

小林正明^{1,2}・門田有希³・大柳 一^{1,2} (1. 明治大・農, 2. JST・CREST, 3. 岡山大・院環境生命科学)

NGS データを扱う際、Web ツールではなくコマンドラインを用いることで、より迅速な解析ができる。そこで、本演題ではコマンドラインを用いた NGS データの解析方法を解説、NIG スパコン上での実習を行った。以下、その内容および手順を示す。なお、ここに示す解説は本演題で行った実習内容を基本的にすべて含んでいる。また、必要なファイル・情報は明治大学の Web サイト (<http://bioinf.mind.meiji.ac.jp/kyusyu/>) にまとめている。当日参加できなかった方々も以下の内容と上述の URL を参考にすれば当日と同様のデータ解析実習が可能である。なお、以下の本文中で「ユーザー名」と記載されている箇所は、自身のユーザー名 (kaminuma など) に置き換えて読み進めて欲しい。

1. NIG スパコンのゲートウェイ (gw または gw2) へのログイン

Windows ユーザーは、使用パソコンに TeraTerm をインストールする。ダウンロードサイト (表 2) に接続し、ダウンロードおよびインストールを完了する。TeraTerm を起動し、「ホスト」に「gw.ddbj.nig.ac.jp」または「gw2.ddbj.nig.ac.jp」と入力し、「OK」をクリックする。初めて TeraTerm に接続した際には「セキュリティ警告」が表示されるが、そのまま「続行」をクリックする。「ユーザー名」と「パスワード」に自身のユーザー名とパスワードを入力し、「OK」をクリックする。ログインに成功すると、コマンドラインを入力可能な画面が表示される。以下の操作は、この画面にて行う。

Mac ユーザーに関しては、「アプリケーション」>「ユーティリティ」>「ターミナル」を起動することで、NIG スパコンへログインする。なお、このログイン法に関しては上述を参照されたい。

表 2. NGS 解析用ツール等の URL 一覧

No.	タイトル	URL
1	実習用 web サイト	http://bioinf.mind.meiji.ac.jp/kyusyu/
2	TeraTerm	http://www.vector.co.jp/soft/dl/win95/net/se320973.html
3	bwa	http://bio-bwa.sourceforge.net/
4	samtools	http://www.htslib.org/
5	SAM ALIGNMENT FORMAT	http://bio-bwa.sourceforge.net/bwa.shtml#4
6	WinSCP	http://winscp.net/eng/docs/lang:jp
7	Java	https://java.com/ja/download/
8	IGV	http://www.broadinstitute.org/igv/
9	IGV User Guide	http://www.broadinstitute.org/igv/UserGuide
10	Velvet	https://www.ebi.ac.uk/~zerbino/velvet/
11	Cufflinks	http://cole-trapnell-lab.github.io/cufflinks/

2. ジョブ投入用のノードへの移動

ssh で遺伝研スパコンへログインした直後は、スパコンの入口計算機にログインした状態である。ジョブ投入用のノードにログインするため「`qlogin`」と入力し、Enter キーを押す。パスワードを入力し、再度 Enter キーを押す。

3. データ・ファイルの用意

以下の手順により実習用データ・ファイルを手し、実習に取り組む

1) データのダウンロード

「`wget http://bioinf.mind.meiji.ac.jp/kyusyu/kyusyu.tar.gz`」と入力し、Enter を押す。ファイルのダウンロードには `wget` コマンドを用いる。ダウンロードできれば、`ls` コマンドで `kyusyu.tar.gz` というデータ・ファイルが確認できる。

2) データの解凍

1) によりダウンロードされたデータ・ファイルを解凍する。「`tar xzvf kyusyu.tar.gz`」と入力し、Enter を押す。解凍後には、`ls` コマンドで `kyusyu` というディレクトリが確認できる。

3) データの確認

「`ls -lh kyusyu`」と入力し、Enter を押すと、`kyusyu` ディレクトリに含まれる 7 つのデータ・ファイル（`IRGSP-1.0_genome_chr01_1-1000020.fasta`, `assemble.sh`, `assemble_results`, `map.sh`, `map_results`, `read1.fastq`, `read2.fastq`）が確認できる。

4. マッピング

1) ディレクトリの移動

「`cd kyusyu`」と入力し、Enter を押す。それにより、`kyusyu` ディレクトリに移動する。

2) シェルスクリプトの確認

「`cat map.sh`」と入力し、Enter を押す。それにより、シェルスクリプト `map.sh` の内容を確認できる。`map.sh` スクリプトには、BWA（Li and Durbin 2009）というマッピングツールを用いてペアエンド・リード（`read1.fastq`,

`read2.fastq`, イネのゲノム DNA 配列）をイネのリファレンスゲノム配列（`IRGSP-1.0_genome_chr01_1-1000020.fasta`）にマッピングし、SAMtools（Li *et al.* 2009）を使ってその結果を整理するための命令が記されている。なお、BWA および SAMtools の詳細はそれぞれの Web サイトを参照されたい（表 2）。

3) シェルをキューに登録

「`qsub map.sh`」と入力し、Enter を押す。「Your job XXXXXX ("map.sh") has been submitted」と表示され、ジョブが開始される。

4) キューの確認

「`qstat`」と入力し Enter を押すことで、作動中のジョブを確認できる。「`map.sh`」が表示されなくなれば、ジョブが終了したことになる。

5) マッピング結果の確認

「`ls -lh`」と入力し Enter を押すと、マッピング後に生成された複数のファイルが確認できる。「`less -S map.sam`」と入力し `sam` ファイルを確認する。`sam` ファイルにはマッピング内容が記されている（詳細は表 2 の SAM ALIGNMENT FORMAT を参照）。「`samtools tv view map_sort.bam IRGSP-1.0_genome_chr01_1-1000020.fasta`」と入力すると、マッピング結果を可視化できる。また、「`g`」キーをタイプ後、「`chr01:1000`」と入力すると 1000 bp 付近へ移動できる。

5. IGV を用いたマッピング結果の可視化

IGV は Integrative Genomics Viewer の略称であり、Broad Institute が公開しているマッピング可視化用ツールである（Robinson *et al.* 2011, Thorvaldsdóttir *et al.* 2013）。Windows, Mac および Linux 環境下で作動し、さまざまな機能が付随されているため一般的に広く用いられている。また、本実習では Windows および Mac などローカル・マシンでマッピング結果を可視化するため、NIG スパコンからローカル・マシンへのデータ転送も行った。

1) マッピング結果のダウンロード

・ Windows ユーザー対象

NIG スパコンからローカル・マシンへデータを転送するため、マニュアルに従い、ファイル転送ソフトウェア WinSCP (表 2) のダウンロードおよびインストールを行う。インストールした後 WinSCP を起動し、ホスト名には「gw.ddbj.nig.ac.jp」または「gw2.ddbj.nig.ac.jp」、また自身のユーザー名を入力する。パスワードを入力すると、ローカルおよびリモート（ここではログインしている NIG のスパコンアカウント）のファイルリストが別々の画面に表示される。リモート画面にあるファイル（ディレクトリ）をドラッグし、ローカル画面へドロップすることでファイル（ディレクトリ）の転送が可能である。本実習では map_results ディレクトリを転送した。

・Mac ユーザー対象

「ターミナル」を起動し、「sftp ユーザー名@gw.ddbj.nig.ac.jp」と入力することでリモートアクセスを行う。パスワードを聞かれるので入力する。NIG スパコンアカウントにログインした後、「get_-R_map_results」と入力しディレクトリを転送する。なお、scp コマンドを使用しての転送も可能であり、その場合は「scp_-r_ユーザー名@gw.ddbj.nig.ac.jp:/home/ユーザー名/kyusyu/map_results」と入力し Enter を押す。

2) Java のダウンロードとインストール

IGV は Java 言語で記述されており、その動作には Java が必要である。そのため、Java の専用 URL (表 2) に接続し、ダウンロードおよびインストールする。

3) IGV のダウンロードとインストール

IGV (表 2) に接続し、ダウンロードおよびインストールを完了する。なお、インストールには、名前および所属等を入力する必要がある。

4) IGV の操作方法

インストール済みの IGV を起動し、まずはリファレンスデータを読み込む。「Genomes」>「Load Genome from File」をクリックし、転送済みの map_results ディレクトリにある IRGSP-1.0_genome_chr01_1-1000020.fasta を選択する。続いてマッピング sam ファイルを読み込む。「File」>「Load from File」をクリックし、map_results ディレクトリにある map_sort.bam ファイルを選択する。これによりマッピングされたリードを確認できる。右上の「-」および「+」をクリックすれば確認したいゲノム領域を縮小および拡大できる。最大限まで拡大すると塩基配列を確認できる。リファレンスゲノムと比較し塩基置換が起きている箇所は、異なる色で塩基が表示される。また、挿入（「I」で表示）および欠失（黒線で表示）なども確認できる。リードの塩基配列のコピー、リード情報（名前、染色体上の位置、クオリティおよび塩基配列など）の確認も可能である。詳細は、IGV サイトの IGV User Guide に記されている (表 2)。

6. de novo アセンブル

NIG スパコンにログインし、コマンドラインにより de novo のアセンブルを行う。以下、その内容および手順を示す。

1) シェルの確認

「cat_assemble.sh」と入力し、シェルスクリプトの内容を確認する。assemble.sh スクリプトには Velvet (Zerbino and Birney 2008) を用いたアセンブルの内容が記されている。Velvet の詳細は表 2 の URL に示されている。

2) シェルをキューに登録

「qsub_assemble.sh」と入力しジョブを登録する。画面には「Your job XXXXXX ("assemble.sh") has been submitted」と表示される。

3) 結果ファイル確認

「ls_-lh_out_k51/」と入力し、結果ファイルを確認する。ここでは、Graph, LastGraph, Log, PreGraph, Roadmaps, Sequences, contigs.fa, stats.txt のデータ・ファイルが生成される。

4) コンティグの統計値確認

「tail_-2_out_k51/Log」と入力し、Log ファイルを確認する。本実習で用いたデータの場合、「Median coverage depth = 21.539821」, 「Final graph has 347 nodes and n50 of 4731, max 39635, total 636872, using 0/304360 reads」と表示される。これは、配列数が 347, n50 値が 4,731 bp, 配列の最大長が 39,635 bp および総塩基長が 636,872 bp であることを示す。

7. SSH 接続の終了

SSH 接続を終了するには、「exit」コマンドを実行しジョブ投入用ノードからログアウトした後、再度「exit」コマンドを実行し入口計算機（ノード）からログアウトする。

おわりに

今回のワークショップでは、実際に持ち込んだノート PC で操作しながら、解説と実習も含めて開催することができた。ワークショップを開催する上で、電源用テابلタップの準備など、多くのご配慮をいただいた。本ワークショップは、科研費・新学術領域研究「ゲノム・遺伝子相関：新しい遺伝学分野の創成」との共催、JST CREST「二酸化炭素資源化を目指した植物の物質生産力強化と生産物活用のための基盤技術の創出」との協賛によって実施することができ、多くの参加者にますます重要となる高速シーケンサー・データ解析について紹介することができた。また、本原稿を取りまとめるにあたって、浅野さとみさん（明治大・農学部）に多大なご協力をいただきました。育種学会員および育種学領域全体の研究の新たな展開のサポートとなれば幸甚である。

引用文献

- Abe, A., S. Kosugi, K. Yoshida, S. Natsume, H. Takagi, H. Kanzaki, H. Matsumura, K. Yoshida, C. Mitsuoka, M. Tamiru *et al.* (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* 30: 174–178.
- Baird, N.A., P.D. Etter, T.S. Atwood, M.C. Currey, A.L. Shiver, Z.A. Lewis, E.U. Selker, W.A. Cresko and E.A. Johnson (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3: e3376.
- Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, C. Li, C. Zhu, T. Lu, Z. Zhang *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42: 961–967.
- Kaminuma, E., J. Mashima, Y. Kodama, T. Gojobori, O. Ogasawara, K. Okubo, T. Takagi and Y. Nakamura (2010) DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Res.* 38: D33–38.
- Kodama, Y., M. Shumway and R. Leinonen (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* 40: D54–56.
- Li, H. and R. Durbin (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.
- Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.
- Nagasaki, H., T. Mochizuki, Y. Kodama, S. Saruhashi, S. Morizaki, H. Sugawara, H. Ohyanagi, N. Kurata, K. Okubo, T. Takagi *et al.* (2013) DDBJ Read Annotation Pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. *DNA Res.* 20: 383–390.
- Ogasawara, O., J. Mashima, Y. Kodama, E. Kaminuma, Y. Nakamura, K. Okubo and T. Takagi (2013) DDBJ new system and service refactoring. *Nucleic Acids Res.* 41: D25–29.
- Peterson, B.K., J.N. Weber, E.H. Kay, H.S. Fisher and H.E. Hoekstra (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One* 7: e37135.
- Robinson, J.T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz and J.P. Mesirov (2011) Integrative genomics viewer. *Nat. Biotechnol.* 29: 24–26.
- Thorvaldsdóttir, H., J.T. Robinson and J.P. Mesirov (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 14: 178–192.
- Zerbino, D.R. and E. Birney (2008) Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18: 821–829.

当日 演題

講演

- 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
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・生命情報）
- Linux マシン使用法
矢野健太郎（明治大・農）
- コマンドラインを用いた NGS 解析
小林正明¹・門田有希²・大柳 一¹（1. 明治大・農, 2. 岡山大・院環境生命科学）