氏　　名	**志　波　　優**
学位(専攻分野の名称)	博　士（バイオサイエンス）
学位記番号	乙　第 897 号
学位授与の日付	平成 26 年 6 月 20 日
学位論文題目	**Establishment of methods for microbial genome analysis using next-generation sequencers and their applications for microbial genomics**
論文審査委員	主査　教　授・農学博士　吉　川　博　文
	教　授・博士(農学)　矢　嶋　俊　介
	教　授・農学博士　千　葉　櫻　拓
	理学博士　板　谷　光　泰*

論文内容の要旨

1995 年にインフルエンザ菌の全ゲノム配列が決定されて以来，様々な微生物のゲノム配列がサンガー法を用いたシーケンサーで解読されてきた。2005 年に従来とは全く異なる原理に基づく，最初の次世代シーケンサーが登場した。サンガー法のキャピラリーシーケンサーと比較して解読塩基（リード）毎の長さは短いが，超並列シーケンスにより圧倒的な解読塩基量を達成し，塩基解読コスト・時間は劇的に低下，研究室単位でのゲノム解析も可能な状況となった。しかしながら，次世代シーケンサーを用いてゲノム解析を行うには，出力される膨大な塩基配列を取り扱うための新たなバイオインフォマティクス解析手法の確立が必要である。

本論文は，第二世代・第三世代を含む次世代シーケンサーを用いた微生物ゲノム解析手法の確立とその応用を目的とした。第 1 章では，多サンプルのバイオインフォマティクス解析を並行して行うための解析パイプラインを新規構築し，枯草菌のリシーケンス解析へ応用した。第 2 章では，リシーケンス解析で検出された変異のアノテーションを行うソフトウェアを新規開発し，出芽酵母のリシーケンス解析へ応用した。第 3 章では，新規ゲノム配列決定とゲノムアノテーション手法を確立し，有用乳酸菌の新規完全ゲノム配列の決定へ応用した。

1.　解析パイプライン NSAP の開発と枯草菌リシーケンス解析への応用

次世代シーケンサーの分野は急速に発展しており，そのデータ解析用のソフトウェアは有償・無償を問わず数百種類以上が利用できる。しかしながら，現状では用途別に様々なソフトウェアを組み合わせる必要があり，解析の煩雑さが多サンプル解析のボトルネックであった。そこで本章では，最も実績のある 8 個の解析ソフトウェアを組み合せて実行可能なパイプライン，NODAI Sequence Annotation Pipeline（NSAP）を開発した。NSAP は Ruby 言語で記述されており，解析内容を指示するレシピファイルを作成することで，リード配列の品質管理，マッピング，*de novo* アセンブル，構造変異検出用の様々なソフトウェアを任意に選択し，一括で実行できる。これにより解析時間の劇的な短縮が可能となった。本章では開発した NSAP を枯草菌 168 株のリシーケンス解析へ応用した。

リシーケンス解析の対象としたのは，日本の 9 研究室から集めた 11 の枯草菌 168 株の野生株である。これらの株は元々 1990 年代に国際コンソーシアムでのゲノムプロジェクトが始まった際に，奈良先端科学技術大学院大学の小笠原研究室から各研究室に分譲された。各研究室では独自に株を管理，維持しているため固有変異の存在が考えられた。各株から抽出した DNA からシークエンスの鋳型調製を行い，第二世代シーケンサー Illumina GAII を用いて 75 または 91 塩基長のペアエンドシーケンスを行った。得られたシーケンスファイルを入力として，NSAP によりマッピング，変異検出，*de novo* アセンブリ，構造変異（大規模欠失・逆位）検出の各種ソフトウェア実行を自動で行った。また，検出された変異を第 2 章で開発した変異アノテーションソフトウェア COVA でアノテーションした。

最初に各株の SNP（一塩基多型），Indel（挿入・欠失），大規模欠失の変異の分布を比較した。EU のコンソーシアムから日本に分譲され，日本の分譲元の株であ

* 慶應義塾大学先端生命科学研究所　教授

る奈良株はリファレンス配列に対して 14 変異が存在し，かつ全株で共通に見られる変異であった。これら 14 変異はキャピラリーシーケンスでも確認し，奈良株に固有の変異ではなく，むしろリファレンス配列のシークエンスの誤りであると考えられた。続いて，168 株の研究室間における多様性を評価した。分譲元の奈良株と比較して，各研究室株にはそれぞれ 1 から 7 箇所の固有変異が存在した。これらの変異には 2 つの遺伝子間領域の変異，5 つの同義置換変異，3 つのフレームシフト変異，8 つのミスセンス変異が含まれていた。この結果は，今後のバクテリアのポストゲノム解析における親株のリシーケンス解析の重要性を如実に示すものであり，かつ多サンプルのリシーケンス解析における NSAP の有用性を示すものであった。

2.　変異アノテーションソフトウェアの開発と出芽酵母リシーケンス解析への応用

リシーケンス解析で検出された各種変異のゲノムへの影響を解析するには，アミノ酸置換への影響といった変異のアノテーションが必要である。既存のソフトウェアは主にヒトやマウスなどの高等真核生物用であり，微生物を対象としたソフトウェアがなかったため，新たなソフトウェア COVA（comparison of variants and functional annotation）を開発した。COVA はリシーケンス解析で検出された各種変異（一塩基置換，挿入・欠失，構造変異）のアノテーションを行うだけでなく，サンプル間での変異比較が可能であり，サプレッサー解析等における有効変異の絞り込みに有用である。また，これまで種々の解析ソフトウェアごとに異なる形式のファイルがアウトプットされてくる問題があったが，本ソフトウェアの開発により複数のプログラムの解析結果を単一の形式で閲覧可能になった。COVA は以下のサイトで公開している（http://sourceforge.net/projects/cova/）。本章では開発した COVA を出芽酵母のリシーケンス解析へ応用した。

リシーケンス解析の対象としたのは，一般的な変異処理剤であるエチルメタンスルフォン酸（EMS）と，近年注目されている不均衡変異導入法でそれぞれ変異処理した出芽酵母である。不均衡変異導入法は，DNA 複製時にラギング鎖特異的に変異を導入する手法であり，従来の変異剤処理では得られにくい形質を得ている実績がある。両変異処理株をシングルコロニー化し，5 コロニーずつから DNA を抽出しシーケンスの鋳型調製後に Illumina GAII を用いて 91 塩基長のペアエンドシーケンスを行った。得られたシーケンスファイル，出芽酵母 S288C 株のリファレンス配列を入力して，第 1 章で開発した NSAP により変異を検出し，検出された変異を本章で開発した COVA でアノテーションし，変異の特徴を比較した。

変異処理株に導入された変異を検出するには，最初に親株からリファレンス配列との相違を引き算する必要がある。COVA では変異株と一緒に親株の変異リストを入力することで，最初に親株変異を差し引き，次いで閾値による変異のフィルタリング，変異のスペクトル分類，変異の位置解析，最後に遺伝子への影響解析（アミノ酸置換等）を行う。従来はデータベース，Excel，配列解析ソフトを組み合わて非常に煩雑な作業を行っていたが，COVA により自動化され効率良く変異アノテーションを行うことができた。

両変異処理株の変異スペクトルを比較すると，EMS 処理株では 97% の変異がトランジション変異で，かつ G/C→A/T の変異に偏っていた。一方，不均衡変異導入株では 72% がトランスバージョン変異であり，その塩基置換に偏りはなかった。不均衡変異の有効性は EMS 等の従来手法と比較して，効率良く多様な変異を導入できる点にあることが示唆され，かつリシーケンス解析における COVA の有用性が示された。

3.　新規ゲノム配列決定の手法確立と乳酸菌ゲノム配列の決定

次世代シーケンサーにより，微生物のドラフトゲノムは容易に得られるようになった。ドラフトゲノムは数十から数百のコンティグ/スキャフォールドで構成されているが，rRNA オペロン，ファージ領域，IS 等のリピート配列を含む領域は部分配列しか得られず，ゲノム構造全体の解析を行うには不十分である。Illumina や 454 等の第二世代シーケンサーのみでは，ドラフトゲノムのフィニッシング作業は依然として労力を要するが，リード長が飛躍的に向上した第三世代シーケンサーによりフィニッシング作業無しで完全ゲノム配列が得られる可能性も出てきた。本章では，第三世代シーケンサーである PacBio RS を用いた新規ゲノム配列決定とゲノムアノテーション手法を確立し，乳酸菌の新規完全ゲノム配列の決定へ応用した。

対象とした腸球菌 *Enterococcus mundtii* QU 25 株は，グルコースやキシロースから L-乳酸を高効率で生産できる菌株であり，完全ゲノム配列を得ることでその性質の特徴に遺伝子だけでなくゲノム構造の観点からも迫り，育種へ応用することが可能となる。最初に，第二世代ショートリード型の Illumina GAII を用いて，ショー

トインサートと 8kb のメイトペアライブラリを構築して *de novo* アセンブルを試みた。塩基長 100bp のリード配列から 310 個のコンティグが得られた。続いて，第二世代ロングリード型の Roche GS FLX +を用いて，フラグメントライブラリのシークエンスを行った。平均塩基長 455bp のリード配列からは 60 個のコンティグが得られたが，依然としてフィニッシングには手間を要すると考えられた。

そこで，第三世代シーケンサーで超ロングリード型の PacBio RS でのシーケンスを試みた。低精度のロングリードの誤りを高精度のショートリードで補正することで，平均 3.7kb，最大 20kb の配列が得られた。7kb 以上の配列を用いてアセンブルを行った結果，約 3Mb のコンティグ 1 本と数十から数百 kb のプラスミドコンティグが 4 本得られた。*Nco*I 制限酵素の全ゲノムオプティカルマップを用いてアセンブルを検証した結果，3 Mb のコンティグにアセンブリの誤りは認められなかった。また，高精度な Illumina のショートリードを PacBio で得られたコンティグにマッピングし，相違は 100 箇所程度であったことから，PacBio RS 単独で得られた配列が極めて高精度であることが明らかとなり，フィニッシング作業は不要であった。

PacBio で得られた配列のアセンブル結果に 3kb の最小プラスミド配列の情報（第二世代シーケンサー結果）を加えた結果，QU 25 株のゲノムは，3.2Mb の環状染色体，5 個の環状プラスミドから構成されていた。染色体上には 2,900 個の ORF，63 個の tRNA，6 個の rRNA オペロンがコードされていることが予測された。また，3 カ所のプロファージ領域，33 個の IS が検出された。また，キシロースからの乳酸発酵に関わる遺伝子としては，22kb の領域に 2 つの遺伝子クラスターが見出された。さらに，QU 25 株の乳酸発酵の特徴を探るために，KEGG Pathway 解析を行った。代謝経路毎の KO（KEGG Orthology）遺伝子数を近縁 *Enterococcus* 属 6 種と比較した結果，QU25 株は ABC transporter，Phosphotransferase system （PTS） の遺伝子数が近縁種よりも多く，高い糖輸送能を持つことが示唆された。以上，本章では第三世代シーケンサーを用いて高精度な完全ゲノム配列を得られることを実証し，QU 25 株のゲノム情報基盤を構築することができた。

まとめ

本論文では，次世代シーケンサーを用いた微生物ゲノム解析手法を確立し，その応用を試みた。第 1 章では，効率的な多サンプルゲノム解析パイプライン NSAP を開発し，枯草菌のリシーケンス解析に応用した。解析時間が劇的に短縮されただけでなく，パラメータ最適化のための反復解析が可能になり，解析精度の向上に貢献した。第 2 章では，リシーケンス解析で検出された各種変異のアノテーションと比較を行う COVA を開発し，出芽酵母のリシーケンス解析に応用した。COVA により各種変異のゲノムへの影響を迅速に解析することが可能となり，変異の比較機能を親株と変異株間の比較に適用することで，有効変異の絞り込みが容易となった。本論文の第 3 章では，主に第三世代シーケンサーを用いて，新規ゲノム配列決定とゲノムアノテーション手法を確立し，腸球菌の新規完全ゲノム配列の決定へ応用した。第二世代シーケンサーでは不可能であったフィニッシング作業無しでの完全ゲノム配列の取得が，第三世代シーケンサーの進歩により可能となったことを，*E. mundtii* QU 25 株のゲノム解析により実証した。QU 25 株のゲノム情報基盤の構築により，今後 QU 25 株の分子生物学的改良への道を拓いた。以上，本論文で確立した解析手法が，今後の微生物学の発展だけでなく，有用微生物の育種等の産業応用にも貢献できると期待される。

審査報告概要

平成 26 年 4 月 23 日（水）午後 4 時から，本専攻が 11 号館 2 階バイオサイエンス専攻大講義室にて開催した学位請求論文の公開本人口頭発表会で，学位請求者志波優氏は，40 分間の口頭発表を行い，その後 20 分間の質疑応答を受けた。発表会終了後，主査，副査と専攻委員による審査会議を開催し，提出論文の内容と本人発表ならびに質疑応答について慎重に審査した。その結果，学位請求者の経歴や学術業績が学位記申請の要項を満たしており，質疑に対する応答が適切だと判断された。また，公表論文に関与した共同研究者との間で学位取得に関して問題が無いことを確認し，当該学位請求論文の内容が学位授与に相当することを全員一致で評決した。

よって，審査員一同は博士（バイオサイエンス）の学位を授与する価値があると判断した。