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 ２８ 日現在

機関番号：８２５０８
研究種目：基盤研究(C)（一般）
研究期間：2012～2015
課題番号：２４５１０２８６

研究課題名（和文）遺伝子機能に影響を及ぼしうる一塩基多型の同定

研究課題名（英文）Infererence of the SNP effects on gene functions

研究代表者
平川　英樹（Hirakawa, Hideki）

公益財団法人かずさＤＮＡ研究所・技術開発研究部・クループ長

研究者番号：８０３７２７４６

交付決定額（研究期間全体）：（直接経費）　　4,400,000円

研究成果の概要（和文）：NCBIのdbESTのEST配列とSRAのシークエンサー由来の配列（SOLiDやHiSeqなど）に対応したSNP解析パイプラインを構築した。トマト品種間でゲノムワイドなSNPを検出しSNPが遺伝子機能に及ぼす影響を推定した。タンパク質立体構造をモデリングし活性部位に位置するSNPを「機能情報をもつSNP」として推定した。病害抵抗性やストレス耐性、果実の色や芳香などに関連する原因遺伝子のうちSNPをもつものを調べた結果、糖代謝などに関連するタンパク質に「機能情報をもつSNP」が見られた。今後、公共データにおける配列データが蓄積し表現型データが充実することで育種効率が向上されることが期待される。

研究成果の概要（英文）：The SNP analysis pipeline for EST and NGS reads sequenced by ABI SOLiD, Illumina HiSeq and MiSeq platforms was constructed. The sequence data were respectively obtained from NCBI's dbEST and SRA databases. Using the pipeline, SNPs on the tomato genome sequence were detected and classified into eight categories, and the effects of the SNPs on protein functions were inferred. The SNPs on the active site in the protein three-dimensional structure constructed by homology modeling were defined as 'functional SNP'. SNPs on the genes related to disease resistances, stress tolerances and fruit color and aroma etc were investigated, and some SNPs were found in the genes related to sugar metabolism and so on. In future, the amount of data related to NGS, phenotypes, QTLs, and three-dimensional structures of protein would be increased in public database. By using these data, the accuracy of variety identification would be improved, and the increase of the breeding efficiency would be expected.

研究分野：バイオインフォマティクス

キーワード：SNP解析　バイオインフォマティクス　データベース

１．研究開始当初の背景

次世代シークエンサー（NGS）が登場して以来、多様な生物種のゲノム解読に用いられており、植物では 2009 年にキュウリのゲノム配列が解読され、その後、トウモロコシやリンゴ、ジャガイモ、そして、2012 年にはトマトのゲノム配列が解読された。NCBI の公共データベース GenBank における dbEST では EST 配列、nuccore では遺伝子配列、SRA ではゲノム由来や転写産物由来の NGS データが公開されている。

トマトは世界中で食され約 8,000 種もの多様な品種が作られており、また、病害抵抗性や耐ストレスに関する遺伝子など古くから研究が進められており、ナス科のモデル植物として、表現型（形質）とその原因遺伝子に関する研究などが行われている。研究開始当初、dbEST には多様な品種の EST 配列がトマトについて登録されていた。また、高密度なプローブをもつイルミナ社製 GoldenGate やより高密度な Infinium ビーズアレイを用いて遺伝子型のデータが得られていた。ロシュ社製 454 やイルミナ社製の GAIIx といった NGS がトマトの幾つかの品種で用いられており、SRA にも数種の品種についての配列データが登録されている状況であった。また、イルミナ社製 HiSeq や MiSeq が登場したことでより膨大な量の NGS データが得られるようになった。これにより、様々な品種についてリシークエンスが行われ、リファレンス配列との一塩基置換（SNP）や挿入欠失（InDel）をゲノムワイドで調べることができるようになった。

２．研究の目的

本研究では、ゲノム配列が解読され、多様な品種について EST 配列が得られており、形質に関わる原因遺伝子が比較的詳しく調べられているトマトを対象としてゲノムワイドな SNP を検出し、それらを分類した後、SNP が遺伝子機能に与える影響を推定する手法を確立することを目的とする。特に SNP が遺伝子のエキソン領域に存在しており、置換によりアミノ酸配列が変わる非同義置換の場合、その SNP がタンパク質の機能部位に位置するかを調べ、さらにタンパク質の立体構造における活性部位に位置する場合、活性を変化させる可能性がある。そこで、直接的に活性に関係する「機能情報をもつ（機能と関わりが深い）SNP」を明らかにし、機能情報をもつ SNP と表現型（形質）との関連性を調べる。表現型については、病害抵抗性やストレス耐性、着色、矮性、果実の色・硬さ・重み、芳香成分といった既報のものを対象とし、様々な品種において得られた遺伝子型と表現型の間で相関関係が高い遺伝子を推定する。

３．研究の方法

NCBI の dbEST からトマト EST 配列を入手した後、poly-A(T)をトリムし、トマトのゲノム配列 SL2.40 に対して GS reference mapper によりマッピングすることで SNP を検出する。SNP を遺伝子との位置関係に基づき以下の８つのカテゴリーに分類する。mSNP（エキソン内；非同義置換）、nSNP（エキソン内；ナンセンス）、sSNP（エキソン内；同義置換）、duSNP（3 ’UTR）、ruSNP（5 ’UTR）、ijSNP（遺伝子の上下流 2 kb 以内）、gSNP（ゲノム領域）、iSNP（イントロン）。全遺伝子の機能情報を深めるため、NCBI の KOG データベース、KEGG の GENES に対するホモロジー検索、Pfam データベースに対するドメイン検索を実施し機能推定を行う。GO（Gene Ontology）については SL2.40 の ITAG2.3 に記載されている情報を用いる。SNP が活性に直接的に影響を与える可能性を調べるため、全遺伝子に対して Modeller を用いてホモロジーモデリングを行い、FPocket を用いてタンパク質の表面の構造から活性部位を推定する。活性部位に含まれる SNP を「機能情報をもつ SNP」と定義する。

一方、我々は、GoldenGate と Infinium ビーズアレイによりトマト 40 系統について 7,054 箇所の SNP データを得たため、構築した SNP 検出パイプラインを適用することで、SNP を上記８つのカテゴリーに分類し機能に及ぼす影響を調べる。

SNP を検出し分類するパイプラインを構築したが、同時期に、次世代シークエンサーにより得られたリードを用いて SNP 解析を行うプログラムが公表されたため、解析手法を見直した。EST 配列のマッピングには TopHat、マップされた領域の抽出には BEDtools、SNP 抽出には SAMtools を用い、SNP の分類（SNP アノテーション）には SnpEff を用いる。また、dbEST のバージョンを 195 に更新し、トマト 44 系統 294,048 本の EST 配列を対象として SNP 検出を行う。

また、ゲノムワイドな SNP を得るため、NCBI の SRA データベースで公開されている NGS のデータを用いる。AB 社製 SOLiD のゲノム由来の配列データに対するマッピングには LifeScope、SNP 検出には VarScan、SNP アノテーションには SnpEff を用いた解析パイプラインを構築する。また、イルミナ社製 HiSeq や MiSeq に対応したパイプラインを構築する。

４．研究成果

NCBI の dbEST からトマト 55 系統を含む 707,445 本の EST 配列を選出し、poly-A(T)をトリムした。こうして得られた 55 系統の EST 配列をトマトゲノム配列 SL2.40 に対して GS reference mapper を用いてマッピングすることで 18,021 箇所の SNP を検出した。このうち 29 箇所は dbSNP に登録されていたが、残り 18,021 箇所は新規なものであった。さらに、各 SNP を遺伝子との位置関係に基づき以下のカテゴリーに分類した（括弧内は

SNP 数)。mSNP（2,008)、nSNP（41)、sSNP（2,150)、duSNP（399)、ruSNP（323)、ijSNP（91)、gSNP（12,020)、iSNP（1,017)。これら一連の解析をパイプライン化することで、今後のデータ更新に対応できるようにした。

上記の解析をしている間、我々は、ゲノムワイドな SNP を得るため、GoldenGate と Infinium ビーズアレイを用い、トマト 40 系統について 7,054 箇所の高精度な SNP を得た。そこで、構築した解析パイプラインと同様の手法で SNP を分類した。既報である病害抵抗性やストレス耐性、着色、矮性、果実の色・硬さ・重み、芳香成分といった表現型に関連する 70 以上の原因遺伝子における品種間の SNP を調べることで表現型に関連する「機能情報をもつ（機能と関連性が深い）SNP」を推定した。その結果、200 個の遺伝子については、SNP が機能部位に存在していたため、機能性マーカーとして用いることができうると考えられた。「機能情報をもつ SNP」をもつ遺伝子について KOG 解析を行った結果、翻訳後修飾、タンパク質代謝回転、シャペロン（KOG O)、エネルギー生産、変換（KOG C)、炭水化物輸送、代謝（KOG G)、アミノ酸輸送、代謝（KOG E)、二次代謝産物合成、輸送、異化（KOG Q）に関連する遺伝子が多く見られた。特に、トマト遺伝子 Solyc09g010080 は、糖代謝に関与するタンパク質のうち、果実の糖含量に関与するシロイヌナズナの Concanavalin A-like lectins/glucanases （brix9-2-5 ; PDB id: 2AC1）に対して 51%の相同性があり、ホモロジーモデリングを行った結果、推定された活性部位における Asn315 が Asp に置換されており、品種間で活性が変化している可能性が考えられた。このように、これら 200 個の遺伝子は品種間で活性が変化しており、形質の違いに関与している可能性があると考えられた。

上記の解析を行った後、SNP を同様に分類（アノテーション）できる SnpEff やマッピング、SNP 検出のプログラムが開発されたため、解析手法を再検討した。EST 配列のマッピングには TopHat、マップ領域の抽出には BEDtools、SNP 抽出には SAMtools を用い、SNP アノテーションには SnpEff を用いた。また、dbEST のバージョンを 195 に更新し、配列を精査することで、トマト 44 系統、294,048 本の EST 配列を対象として SNP 検出を行った。その結果、145,880 本（49.6%）がマップされ、38,684 箇所の SNP が検出された。SNP の数が最も多かった品種は TA496（加工用;108,440 本のうち 17,654 本（16.3%）がマップ）であり、次いで Micro-Tom（矮性；118,119 本のうち 10,034 本（8.5%））であった。検出された SNP のゲノム上の位置を元に、SnpEff を用いて６種類の SNP（エキソン（同義置換、非同義置換)、イントロン、5’ UTR、3’ UTR、その他）に分類した。マッピングにより得られた vcf ファイルの DP4 情報を元にして、マッピングの精度、厚み、リファレンスと EST 配列の順鎖方向と逆鎖方向の厚みを全系統に対して集計した表を作成した。

こうして EST 配列に対応した解析パイプラインを構築したが、NCBI で公開されている EST 配列では本数が少ないため十分な精度が得られないこと、ゲノムワイドな SNP が得られないこと、品種によって EST 配列の本数に偏りがあるため全ての品種において特定の箇所の塩基置換を比較することが困難であることが問題点として挙げられた。

一方、公共データベース NCBI の SRA データベースで公開されている NGS のデータではリード量が多いためゲノムワイドな SNP が検出できると思われた。そこで、AB 社製 SOLiD により得られたトマト 6 品種（Ailsa Craig、Furikoma、M82、Tomato Chukan Bohon Nou 11 Gou、Ponderosa、Regina）のリードに対して Lifescope によりマッピングし SNP を抽出した結果、リードのマップ率は 75.0～77.9%、リファレンス上のカバー率は 92.7～93.4%となったが、リードの厚みが 8.5～17.3 となり、十分なリード量では無く検出された SNP の精度も十分では無いと考えられた。しかしながら既知の原因遺伝子 78 個のうち 72 個において品種間で SNP が検出され、そのうち 32 個については非同義置換であった。こうして SOLiD と 454 に対応した解析パイプラインを構築した。

この時点でトマトのゲノム配列が SL2.50 に更新されたため、全遺伝子に対する機能推定や KOG による機能分類、KEGG による代謝経路の推定、活性部位の予測といった一連の解析を再度実施した。

イルミナ社製 HiSeq や MiSeq といった NGS が登場し、大量の配列データが得られるようになったため、リシークエンスに用いられることでゲノムワイドな SNP が得られるようになった。そこで、EST 配列について構築したパイプラインを HiSeq と MiSeq に対応させた。その際、ゲノム由来のリードも解析できるよう Bowtie 2 を適用した。本研究で構築した解析パイプラインは Linux のコマンド上で実施するものであり、実行できるユーザが限られていた。そこで、ウェブ上からパイプラインを実行できるインターフェースを開発した。今後、改良を重ね公開する予定である。

現在、NCBI の SRA データベースには多様なトマト品種について 1,500 を超えるエントリーが登録されている。それらの登録内容にはリード長やペアエンド、メイトペアといったリード情報、品種や掛け合わせといったサンプル情報が記載されている。SRA データベースではエントリー情報をテーブル形式で入手できるが、品種情報などの詳細については記載されていないことが多い。そこで、SRA に登録されたエントリーについて、すべ

ての情報を XML 形式で入手し、表に変換するプログラムを開発した。現時点では、500 を超える品種が登録されており、それらについてゲノムワイドな SNP を検出できるようになった。これらの品種の数が多いため現在も実行中であり、期間中に解析を終えることができなかった。このように SRA には大量のデータが公開されているため、今後、解析パイプラインにおける処理を並列化することで効率を上げる必要がある。一方、表現型のデータに関しては、病害抵抗性やストレス耐性などの原因遺伝子が文献を通じで公表されている。今後はこれらの情報をキュレーションし、その結果をデータベース化することも重要である。研究を開始した当初は SRA において品種が記載されたエントリーは 10 程度であったが、現在はリシークエンスにより品種の数が増加している。今後もシークエンス技術の向上によりさらに多くの品種についてのシークエンスが蓄積されていくと考えられる。公共データを用いた SNP 解析や表現型データの充実により遺伝子型と表現型との相関関係の研究を実施することで、「機能情報をもつ SNP」が明らかになると考えられる。これにより品種選抜が効率良く行われ、従来では時間を要していた育種の効率も向上されることが期待される。

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）

Hirakawa H., Shirasawa K., Ohyama A., Fukuoka H., Aoki K., Rothan C., Sato S., Isobe S., Tabata S. Genome-wide SNP genotyping to infer the effects on gene functions in tomato. DNA research. 2013, 20, 221-233.

〔学会発表〕（計 2 件）

①Hirakawa H., Shirasawa K., Isobe S., Sato S., Ohyama A., Fukuoka H., Tabata S. (2011) Functional analysis of genes based on large-scale SNP data in tomato lines. SOL & ICuGI 2011, 8th Solanaceae and 2nd Cucurbitaceae Joint Conference, Kobe, Japan (Oral) 2011.11.28-12.2
②Hirakawa H., Shirasawa K., Fukuoka H., Aoki K., Asamizu E., Sato S., Isobe S., Tabata S. (2013) Genome-wide SNP genotyping of tomatoes using array and NGS data. Japanese Solanaceae Genomics Initiative (JSOL) 10th International Symposium on Solanaceae Genomics, Kyoto, Japan (Oral) 2013.11.29-30

〔図書〕（計 0 件）

〔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：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
平川　英樹（HIRAKAWA, Hideki）
(公財)かずさ DNA 研究所・グループ長
研究者番号：80372746

(2)研究分担者
なし

(3)連携研究者
なし