様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　１８　日現在

機関番号：　８２１１１
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ０７１７５

研究課題名（和文）大規模ゲノム配列情報解析による作物栽培化過程の解明

研究課題名（英文）Crop Cultivation Process Deciphered by Large-Scale Genome Sequence Analysis

研究代表者
伊藤　剛（Itoh, Takeshi）

国立研究開発法人農業・食品産業技術総合研究機構・高度解析センター・チーム長

研究者番号：８０３５６４６９

交付決定額（研究期間全体）：（直接経費）　　4,000,000円

研究成果の概要（和文）：作物栽培化過程の重要な側面として分岐の時間や集団の大きさを明らかにするため、大規模配列解析による方法の構築を行った。これは大量のゲノム配列決定が安価に行えるようになったことにより、可能となった。全ゲノム配列を二つの近縁種や近縁集団から取得し、まず全ゲノムでの置換数の分布を作成する。この分布が、座位間で均一な置換発生を仮定する単純なポワソン分布に合わないことは非常にはっきりしているが、負の二項分布に当てはめるよりむしろ種分岐と祖先多型の両方を考慮したモデルの方がよく当てはまることが分かった。例えばアジアとアフリカのイネでは、従来言われていた40万年よりはるかに小さな値を推定することができた。

研究成果の概要（英文）：In this study, to elucidate the process of ancient crop cultivation, it was aimed that the divergence time and effective population size would be estimated accurately through large-scale analysis, which was recently enabled by low-cost, high-throughput DNA sequencing. Whole genome sequence data are first obtained from two different closely-related species or populations, and then a genome-wide distribution of nucleotide substitutions is calculated. The distribution does not obey a Poisson distribution, which assumes an equal substitution rate. Although a negative binomial distribution better fitsto the data, a mixed model of speciation and ancestral polymorphisms was found to best fit to the data. In the data analysis of this study, for example, the divergence time between Asian rice and African rice was much smaller than the previous estimate, which means that the past estimation overlooked the inflation of the estimated values that was caused by ancestral polymorphisms.

研究分野：　分子進化

キーワード：　栽培化　分岐時間　集団の大きさ　大規模ゲノム解析

１．研究開始当初の背景

（１）栽培化の過程は、考古学的な遺物等から推定することができる。しかし、植物ははっきりした痕跡を残さないことも多く（Fuller、2007、Annals of Bortany）、直接的な証拠を得ることが難しい。本提案では、大量の分子データから間接的に、しかし確実な情報を得ることを考える。特に、品種間あるいは品種と野生種間の分岐パターンと分岐年代を推定し、栽培作物の由来を明らかにする。

（２）近年、いわゆる次世代シーケンサー（NGS）による DNA 解読技術の革新が続き、大量の塩基配列を低コストで読み取ることが可能となった。まず、多くの作物で参照ゲノムや大量完全長 cDNA が解読されており、これがデータ解析の基盤となる。公的に利用可能なデータが豊富に存在している一方、多くの研究の NGS データでは種や品種あたりのデータ量が少ない。例えば野生イネのゲノム解読（X. Huang ら、2012 年、Nature）では非常に多数のイネ野生種を対象としているが、サンプルあたり 1～2 倍程度のカバレッジであって、全ゲノムに渡る精密な研究には向かない。このため、興味ある種や品種によっては、改めて厚みを持った大規模解読も必要である。

２．研究の目的

（１）人類にとって「食」はその生存の大きな部分を占め、従って栽培作物の改良は絶えざる興味関心の的である。これからの栽培種の改良を考える上でも、ヒトが行ってきた栽培化の歴史を明らかにし、いつどのようにして作物が形作られたのかを知ることの意義は大きい。本提案では、ヒトによる作物栽培化の歴史を分子進化の手法による徹底的な大規模データ解析から解き明かすことを目指す。

（２）この研究の範囲では、理論構築並びに実データの解析を比較的小規模に一部の作物で行うことで、該当研究の技術的諸問題を克服し、成果を上げることを狙いとしている。ただ、ここで得た成果をもとに、将来的には人類全体の作物栽培史のような、より大きな研究展開につなげていくことも期待される。

３．研究の方法

（１）品種間や近縁種の微細な違いを明らかにするため、高精度の参照ゲノムがあることが望ましい。その意味ではアジアイネが最善である。ジャポニカ（日本晴）に対して、アフリカイネ等のデータががすでに公開されていることから、これを活用する。この他にムギ類、カンキツ類も直ちにデータが活用可能であるので対象とする。

（２）参照ゲノムに対して BWA などのプログラムで NGS のリードをマッピングし、samtools 等で SNP データを作成する。複数種・品種の SNP を参照ゲノムに対して明らかにした場合、塩基座位ごとの樹形計測の方法により、incomplete lineage sorting を考慮しながら系統関係を推定できる。分岐年代等推定のためには、ゲノム配列を小さく分割（1 kb 程度）したウインドウで配列間の違い（置換）を計測し、この分布を作成する。置換数分布は、種分岐と祖先多型の合成したものである。また、種間もしくは集団間の塩基置換は、理想的にはポワソン分布に従うはずであるが、実際には負の二項分布（ポワソン-ガンマ分布）になることが期待される。データから期待値と分散を得た場合はこれらから形状母数を推定できる。

（３）パラメーターの推定には最尤法が多く用いられる。しかし最尤法は計算量が多く、プログラムを並列化してもなお計算時間を要する。そこで、PC クラスターでの分散計算などに対応したプログラムを開発し、高速に計算ができるようにする。

４．研究成果

（１）イネにおいて、提案した通り、種分岐と祖先多型の両方を考慮したモデルを基に作成したプログラム（並列計算が可能）を用いて分岐年代および有効な集団の大きさを推定した。栽培種であるサティヴァのジャポニカの日本晴品種と、新規に次世代シーケンサーの HiSeq でゲノム解読したアフリカイネであるグラベリーマ（IRGC104038）を比較し（図 1）、15～17 万年前という分岐年代、11 万強の集団の大きさを得た。

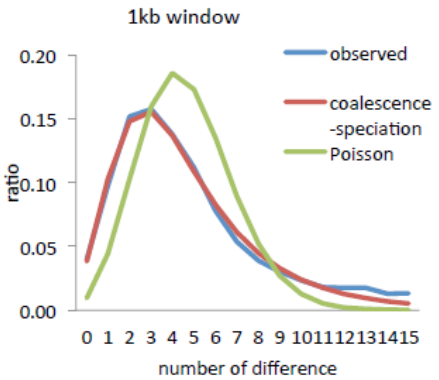

図１　アジアイネとアフリカイネの塩基置換分布。

同様に、ジャポニカの日本晴とインディカの広陸矮 4 号を比較したところ、この場合は最初の集団の分岐後に再度交雑があり、この交雑の時期は非常に現代に近いことが分かった。また、ダイズの栽培種（Glycine max）と野生種（ツルマメ、Glycine soja）にも本方法を適用し（図 2）、分岐年代が数千年前であったことから（表）、栽培化の歴史に関する考え方とも比較的合致し、短期間での集団の分岐や進化に対しても提案の方法が使え

ることを示すことができた。

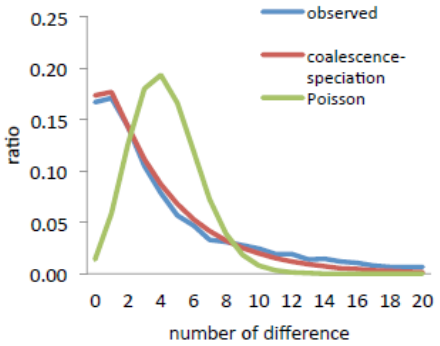

図２　ダイズ種間の塩基置換分布

ウインドウ幅	分岐時間	集団の大きさ
1 kb	2659	91279
2 kb	6203	91944
3 kb	7681	75625

表　ダイズの分岐時間等推定値

これらに加え、新規のゲノムシーケンシングを企図し、多様な種を解析するため、我々が食する温帯ジャポニカとは異なる熱帯ジャポニカの WRC52（ベトナム）、アフリカの野生種であるバルチーから W0698（ギニア）と W1588（カメルーン）、さらにこれまでに述べたイネの外群として Oryza longistaminata の W1449（コートジボワール）を HiSeq にて解読した。加えて、ビグナ属の野生種である Vigna exilis も解読した。これらに関しては、参照ゲノムに対してアラインメントし、比較検討を行った。

（２）参照ゲノムが最もしっかりしており、またアジアを中心として極めて重要な作物であるイネを中心とし、ダイズを加えるなどして解析を行ってきたが、その結果、本提案で考案している分岐年代と有効な集団の大きさを推定する方法は、十分に有効であることを明らかにしてきている。そこで、はさらに対象生物種を広げ、まずムギ類を検討した。塩基配列解読が非常に高速かつ低価格になった現在ではあるが、それでも懸念した通り、６倍体コムギはゲノムが非常に大きく、Illumina HiSeq X Ten などであっても十分な解読量が確保できるかどうか不安が残る。加えて、コムギゲノムの参照配列は残念ながら最終版の公開に至っていない。そこで、参照ゲノムの最終版が公開されており、２倍体であるオオムギを中心に考えることとした。栽培種を２種、それぞれで約 14 億リード（150 bp）を、また近縁野生種で約 13 億リードを得た。低クオリティ領域の除去後に 5%程度のデータ量減少が見られたが、それでも塩基量で全ゲノムの 30 倍程度以上は得られているので、解析に十分なデータが確保できた。また、野菜や果樹も対象として検討していたところ、別途共同で行っているカンキツ類のゲノム解読のデータが使用できるので、野生種のカラタチとウンシュウミカン、ブンタンの配列比較を行うこととした。オオムギとカンキツのデータ解析については、ゲノムとのアラインメントを作成し、全ゲノムでの置換のデータを得た。

（３）種間もしくは集団間の塩基置換は、座位間に置換速度の差がなければポワソン分布に従うはずである。しかし実際には、自然選択その他の要因により、座位ごとに置換速度の差があり、これがガンマ分布に従うと仮定するとポワソン分布より分散が大きい負の二項分布（ポワソン-ガンマ分布）になるはずである。負の二項分布の平均の期待値はデータ全体の平均値と一致し、これは最尤推定値でもある。また、分散は期待値と形状母数（gamma parameter）から計算され、データから期待値と分散を得た場合はこれらから逆に形状母数を推定できる。例えばイネにおいてアジア栽培稲とアフリカイネの間の置換数の分布を作成し、ポワソン分布と比較すると大きくずれる。ここで負の二項分布を仮定して比較すると、ポワソン分布よりはるかによく実データと一致することが分かった。ただ、祖先多型を考慮して、種分岐による部分と祖先の種内多様性による部分を分離した場合、更によく実データと合致するので、二者から択一の場合はこちらの方がよい（図 3）。

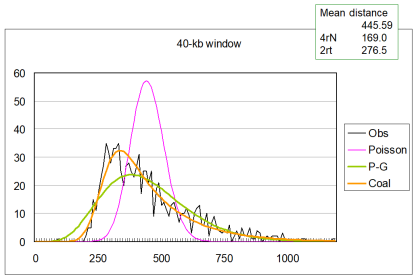

図３　実データ（黒）、ポワソン分布（ピンク）、負の二項分布（緑）、祖先多型＋種分岐（オレンジ）

この後者に負の二項分布を導入すると数式が大変複雑になり、今のところ導入には成功していない。

５．主な発表論文等

なし。

６．研究組織

(1)研究代表者

伊藤　剛（ITOH, Takeshi）

国立研究開発法人農業・食品産業技術総合研究機構・高度解析センター・チーム長

研究者番号：80356469

(2)研究分担者

なし

(3)連携研究者

坂井寛章（SAKAI, Hiroaki）

国立研究開発法人農業・食品産業技術総合研究機構・高度解析センター・主任研究員
研究者番号：20455322

熊谷真彦（KUMAGAI, Masahiko）
国立研究開発法人農業・食品産業技術総合研究機構・高度解析センター・任期付研究員
研究者番号：80738716

(4)研究協力者
なし。