有用遺伝子発掘を目指した日本産ホタルのゲノム解析と遺伝子比較

加藤太一郎

鹿児島大学学術研究院理工学域理学系
〒890-0065 鹿児島市郡元 1-21-35
e-mail : kato@sci.kagoshima-u.ac.jp

要旨

日本にはおよそ 40 種類のホタルが知られているが、鹿児島県島嶼地域にはそのうちの約半数 20 種類に及ぶ陸生・水生ホタルが自生している。また島々にも固有のホタルが生息していることも特筆すべき特徴である。日本に生息する最も有名なホタルの一つであるゲンジボタルは、生息地によって発光特性や産卵様式が違っていることが分かっており、ミトコンドリア上の COI 遺伝子の配列の違いにより大きく 6 つのタイプに分類できる。しかしその表現型の違いが引き起こされる理由については未解明のままであり、謎の解明には全ゲノム比較を行うことが必要である。本研究では、ホタルゲノム全体での遺伝子多様性について議論することを目指した全ゲノム解析を試みたので報告する。研究の結果、ゲンジボタルゲノム DNA は高いヘテロ接合度や、短いイントロン、および低い GC 含量等、既に解析されている他の生物とは異なる特徴的な性質を示すことが明らかになった。ゲノムアノテーションは RNAseq 配列比較法、Ab-initio 法、およびホモロジーサーチ法の 3 つの手法を統合することで行った。本研究期間内に統合を完了することはできなかったが、およそ 15,000 の遺伝子を持つことを確認した。

1. 緒論

1-1 ゲノム解析の進歩

オックスフォード・ナノポア社の提供する MinION は、ポケットサイズのシーケンサデバイスである。既存の技術に比べて格段に使いやすくなったこのシーケンサの進出は、世の中を驚かせた。ナノポア社の台頭によって、今後ゲノムシーケンシングはより我々の身近なものになっていくと予想される。現に、ゲノム創薬や遺伝子性疾患の予防・治療など、医療分野においてゲノム情報は大いに貢献している。

1-2 ホタルのゲノム解析について

現在、真核生物 2,731 種に関してゲノム解析が完了している。その中で昆虫 *Coleoptera*（甲虫目）は 13 種しか含まれず、*Lampyridae*（ホタル科）は一種もゲノムが読まれていない。本研究では、ホタルに関する分子生物学的、集団遺伝学的知見を広めるため、次世代シーケンス(NGS)技術を用いた DNA 解析を試みることとした。対象は、日本の代表的なホタルである、ゲンジボタルとヘイケボタルを選択し、これらに対してゲノムシーケンス解析および RNAseq 解析を行った。先に述べたように、リファレンスとなる参照ゲノムがないため、すべて de novo によって解析を進めた。

1-3 ホタルのミトコンドリアゲノムアノテーションについて

ゲンジボタルのミトコンドリアゲノムについては、2013 年に松井らが Genbank 登録を行っている（AB849456）[1]。しかし、学術雑誌での報告はなされていなかったため本解析で得られたゲンジボタルとヘイケボタルについてのミトコンドリアゲノムのアノテーションを行った。

2. 解析方法

2-1 ゲノム解析の流れ

ホタルサンプルは奈良県のホタル養殖業者（大和の国かわぐち）から購入したゲンジボタルとヘイケボタルを用いた。ゲノム解析と RNAseq 解析の流れを図 1 に示す。ゲノム解析ではホタルの全身を、RNAseq 解析ではホタルの尾部を用いて、QIAGEN の Blood & Cell DNA Min Kit によってゲノム DNA の抽出を行った。シーケンシングはイルミナ社の Hiseq によって行った。得られたシーケンスリードを基に、ゲノムシーケンスは platanus、RNAseq では DDBJ Read Annotation Pipeline と TRINITY というアセンブラーを使ってアセンブルを行い scaffold の構築を行った。この、scafold の配列に対してアノテーションを行った。シーケンシング、アセンブルは国立遺伝学研究所と共同で行った。RNAseq 解析は DDBJ Read Annotation Pipeline/ Galaxy/P-GALAXY を利用した。ゲノムアノテーションは、三種のアノテーション手法の結果を手動にて統合した。

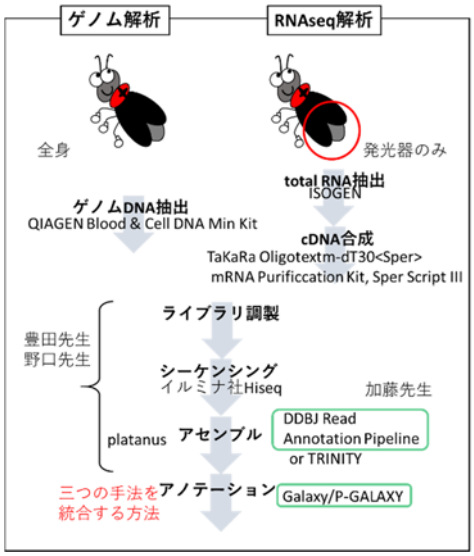

図 1. ホタルゲノム解析・RNAseq 解析の大まかな流れ

2-2 ゲノムアノテーションの方法

ゲノムのほとんどは非遺伝子領域から構成されており、遺伝子領域は全体の 30%、さらにコード領域は全体の 3%と言われている（図 2）。遺伝子領域のアノテーションを行うにあたって、まずは反復配列のマスクを行った。反復配列は、RepBase という反復配列データベースにその情報が記録されているので、これを基に Repeat Masker というソフトによってマスクを行った。次に、細菌のコンタミを除くために、NCBI の 16srRNA データベースを用いて BLAST 検索により細菌の配列を検索し、これを除去した。

上記クリーニングした塩基配列に対してゲノムアノテーションを行った。ゲノムアノテーションは三つの手法を用いて行い、それらのアノテーション結果を評価、統合した。三つの手法の一つ目は、「RNAseq 配列との比較」である。RNAseq によって得られた配列はその生物の転写産物であるので、これを用いてゲノム上の遺伝子領域を調べることができる。二つ目は「ホモロジーサーチ」である。RNAseq 配列との比較は、同じ生物由来の配列を用いるためエビデンスがしっかりしているが、ゲノム抽出を行ったときに転写されていた転写産物について

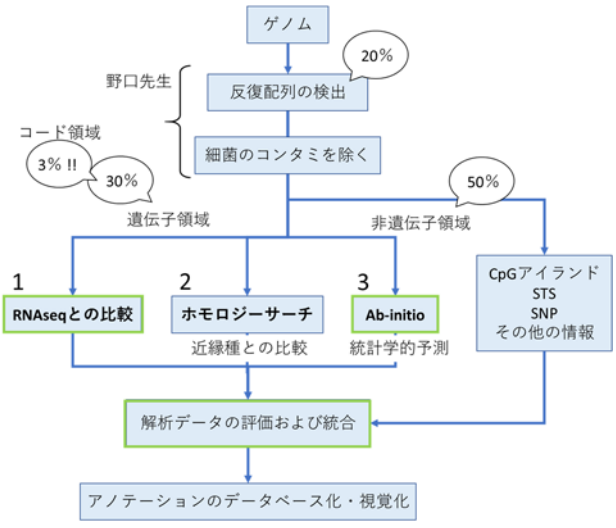

図 2. ホタルゲノムアノテーションの大まかな流れ

しか調べることができないという難点がある。そこで、近縁種のタンパク質データベースを用いて、RNAseq の配列でカバーできなかった遺伝子領域をアノテーションする。三つめは「Ab-initio 法」である。これは、統計学的に遺伝子領域を予測する方法である。RNAseq 配列との比較やホモロジーサーチで拾いきれなかった遺伝子領域まで予測できる可能性もあるが、逆に間違いのアノテーションも多くなってしまう傾向にある。これら三つのアノテーション手法についての特徴を表 1 にまとめる。本研究では、RNAseq 配列との比較と、Ab-initio 法、またアノテーション結果の統合を行うと共に、ホモロジーサーチによるアノテーションは国立遺伝学研究所の協力を仰いだ。

表 1. 三つのアノテーション手法の特徴

	信頼性	完全長CDS配列	感度	特徴
1. RNAseq	○	×	△	partial
2. ホモロジーサーチ	△	△	△	CDSのみの予測
3. Ab-initio	×	△	○	遺伝子を複数繋いでしまう

2-3 ミトコンドリアゲノムアノテーションの方法

ミトコンドリアゲノムは環状であるので、どこを一塩基目とするかをまず決定した。Genbank の他のミトコンドリアアノテーションを見てみると多くは tRNA の Ile から始まっていたのでこれが一塩基目にするように配列を編集した。具体的には BioEdit で tRNA Ile の配列（AB849456）とミトコンドリアゲノムをアライメントして、ヒットしたところが一塩基目になるように、Edit モードで編集した。

次に、CDS 領域と RNA コード領域の決定を行った。AB849456 の CDS 配列・RNA 配列とミトコンドリアゲノム配列を BioEdit でアライメントして CDS の位置を決定した。このとき、ヒットの最初と終わりが開始コドンと終止コドンになっているかを確認した。得られた CDS 配列を翻訳してみて、終止コドンが入らないかを確認した。翻訳するときは、ミトコンドリアゲノム特有のコドンテーブルに変更することに注意する。Genbank では transl_table=5 となっている。通常のコドンテーブルとの違いと、BioEdit でコドンテーブルを変更する方法を表 2 および図 3 に示す。

表 2. ミトコンドリアゲノム特有のコドンテーブル

コドン	cord5	ふつう
AGA	Ser	Arg
AGG	Ser	Arg
AUA	Met	Ile
UGA	Trp	*

コドンテーブルの変更の仕方

↓Option
↓view codon table
↓save（拡張子　.tab）
↓saveしたものをメモ帳で開いて編集

↓Option
↓select con table
↓先ほど編集したコドンテーブルを選択
↓翻訳すると、そのコドンテーブルを用いて翻訳してくれる

図 3 BioEdit でコドンテーブルを変更する方法

一つ目の RNA 配列（tRNA の Ile）と最後の RNA 配列（srRNA）の間を、AT-rich region とした。系統樹の作成は MEGA6 を用いて行い、NJ 法、p-distance で行った。

3. 結果と考察

3-1 ホタルゲノムの特徴

3-1-1 ゲノムの配列情報

アセンブルによって得られたゲノムの配列情報を表に示す（表 3）。アセンブルの指標である N50 値を見てみると、ゲンジボタルが 11.9 Mb、ヘイケボタルが 320.2 Mb となっていた。N50 値は Scaffold（Contig）を長さの長い順につなげていったとき、長さの累積が推定ゲノムサイズの 50%を超えたときの Scaffold（Contig）の長さを表している（図 4）。長い配列が多いと N50 は大きくなり、アセンブルがうまくいっていない短い配列が多いと N50 は大きくなる。復元したいゲノムに少しでも近づけるには長い配列が多く得られた方が良いので、N50 値はアセンブルの結果の良しあしを判断する指標になっている。一般に、N50 値が数 Mb あれば十分に論文化できると言われているが、これと比較するとゲンジボタルのアセンブル状況は非常に良い状態だと言える。逆に、ヘイケボタルのアセンブル状況はあまり良くなく、改善が必要だと判明した。

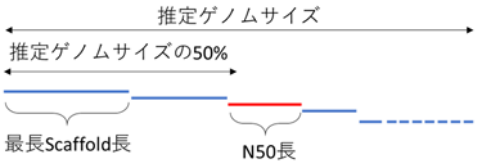

図 4.　N50 とは

表 3. ゲンジとヘイケのゲノム配列情報

ゲンジボタル（Luciola cruciata）推定ゲノムサイズ＝667.9 Mbp

	Min	Num	Total Len	Max	Avg	N_{50}
Contig	1 Kb	7,829	657.0 Mb	1.5 Mb	83.9 Kb	0.3 Mb
Scaffold	1 Kb	3,247	662.9 Mb	45.5 Mb	204.1 Mb	11.9 Mb

ヘイケボタル（Luciola lateralis）推定ゲノムサイズ＝1,022 Mbp

	Min	Num	Total Len	Max	Avg	N_{50}
Contig	1 Kb	87,166	1.37 Gb	0.7 Mb	15.6 Kb	33.7 Kb
Scaffold	1 Kb	20,747	1.58 Gb	9.2 Mb	76.3 Mb	416.8 Kb

	Min	Num	Total Len	Max	Avg	N_{50}
Contig	100	1,778,349	1.72 Gb	0.7 Mb	1.0 Kb	24.2 Kb
Scaffold	100	1,708,424	1.94 Gb	9.2 Mb	1.1 Kb	320.2 Kb

また K-mer による推定ゲノムサイズを見てみると、ゲンジボタルが 667.9 Mbp となっており、Scaffold の Total Len とほぼ同じになっていた。一方、アセンブルがうまくいっていないヘイケボタルでは、推定ゲノムサイズが 1,022 Mbp であるのに対して、Total Len が 1.58 Gb（Min 1kb の場合）、1.94 Gb（Min 100 bp の場合）と約 1.5～2 倍の値を示した。

3-1-2 ヘテロ接合度

ゲノムサイズの推定に K-mer を用いたが、シーケンスリードをある決まった文字数切り出したものである。K-mer はゲノムサイズの推定の他に、アセンブルの際のパラメータやヘテロ接合度の計算などに利用される。例として、ヘテロ接合度を表す、K-mer の頻度分布を下に

示す（図 5）。横軸が K-mer の出現回数、縦軸が K-mer の種類の数を示す。青い線が 1 倍体のゲノムを、オレンジの線が 2 倍体ゲノムでヘテロ接合度が 0.1%の頻度分布を示す。1 倍体ゲノムでは山が一つであるのに対して、2 倍体ゲノムでは大きな山の K-mer の出現回数のちょうど半分の出現回数のところにもう一つ小さな山ができる。小さな山はヘテロ接合になっている配列から得られた K-mer であり、2 倍体であるのでこれの出現回数は大きな山の半分になる。

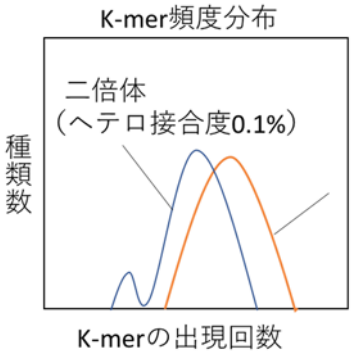

図 5. K-mer の頻度分布の例

次に、本解析によって得られたゲンジとヘイケの K-mer の頻度分布を示す（図 6）。今回は、K=31 とした。赤い線がゲンジ、緑の線がヘイケを表している。先ほどの、ヘテロ接合度が 0.1%程度だと、ヘテロ接合由来の K-mer の種類数が少ないので山が小さくなるが、この結果では、ヘテロ接合由来の 31-mer の種類数がホモ接合由来の 31-mer の種類数を大きく上回っていることが分かる。このような例は、非常に稀であるようだ。先に、ヘイケのアセンブルがうまくできておらず、Scaffold の Total Len が推定ゲノムサイズの 2 倍になっていることを示したが、これはこの異様に高いヘテロ接合度が原因になっていると考えられた。本来、シーケンシングによって得られた配列は、父方と母方の配列がマージされ、1 倍体の形で出力されるが、ヘイケでは父方母方の違いが大きすぎたため、マージすることができなかったのだと予想した。

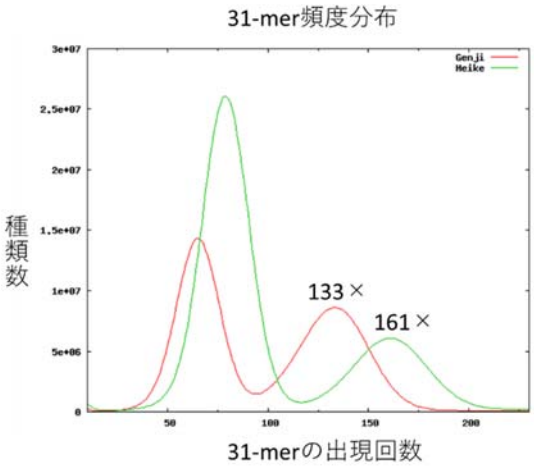

図 6. ゲンジ・ヘイケの 31-mer の頻度分布

3-1-3 ゲンジボタルゲノムのその他の特徴

その他のゲノムの特徴として、イントロンが短いということが判明した。ヒトやショウジョウバエのイントロンのピークが 80 bp ほどであるのに対して、ゲンジではイントロンのピークが 50 bp であった。このことにより、BLAST 検索の際にイントロンを indel と誤認識し、一つのヒットの中に二つのエクソンを含めてしまう傾向があった。

また、GC 含量が他の生物種に比べて低いということも分かった。同じ甲虫目である *T. castaneum* は 35%の GC 含量であるのに対して、ゲンジボタルでは 28.55 %と AT リッチであった。

3-2 ゲノムアノテーション結果

三つのアノテーションの結果を示す。RNAseq 配列との比較によるアノテーションでは 30,926 遺伝子、ホモロジーサーチによるアノテーションでは 17,901 遺伝子、Ab-initio によるアノテーションでは 15,869 遺伝子がアノテートされた。このアノテーション結果についての統合は現在進めているところであるが、最終的に 15,000～20,000 程の遺伝子数になると予想される。近縁種である *T. castaneum* の遺伝子数は 22,611 であり、ほぼ同等の遺伝子数であるとわかった。

3-3 ミトコンドリアゲノムアノテーション結果

アノテーションの結果、ゲンジもヘイケも後生動物と同様の遺伝子構成であった。（図 7）。GENETYX を用いてゲノム構造を可視化した図を示す（図 8）。

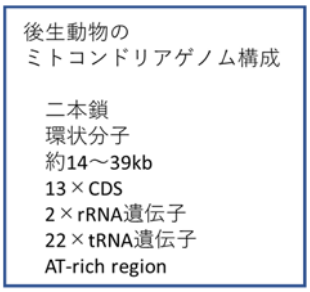

図 7. 後生動物の
ミトコンドリアゲノム構成

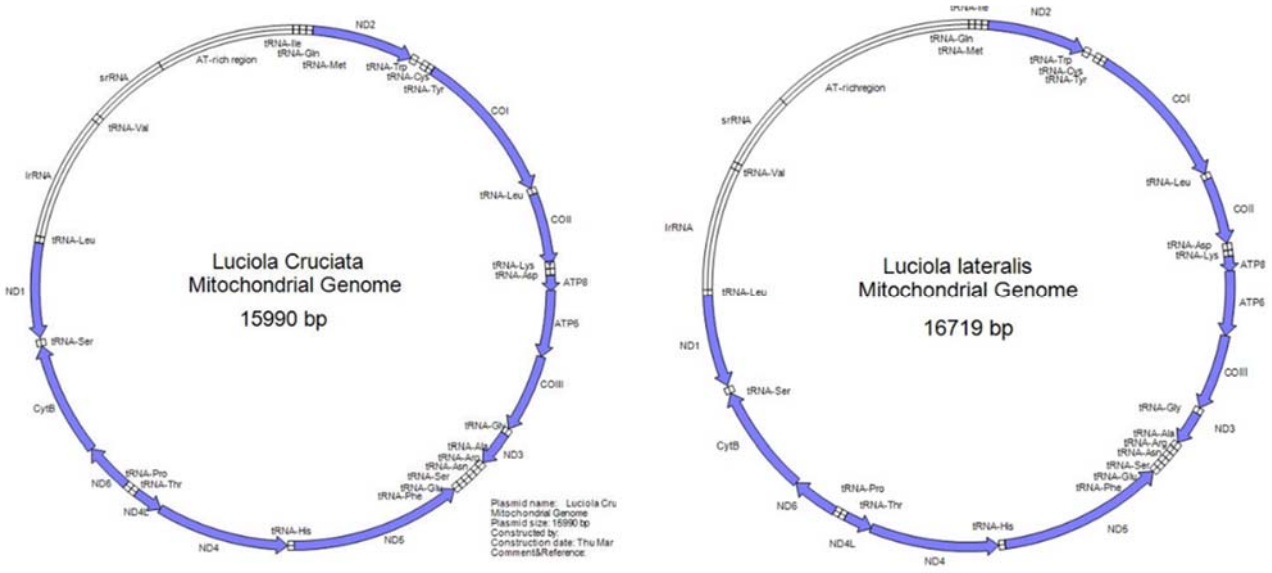

図 8. ゲンジ、ヘイケのミトコンドリアゲノム構成

まず本解析に用いたゲンジボタルと AB84945 を比較すると、それぞれ総塩基数が 15,990 bp と 15,989 bp となっており、本解析のサンプルの方が 1 塩基多くなっていた。二つの配列を BioEdit でアライメントしてみると、AT-rich region の 14,800 塩基目に C が挿入されていることが分かった（図 9）。

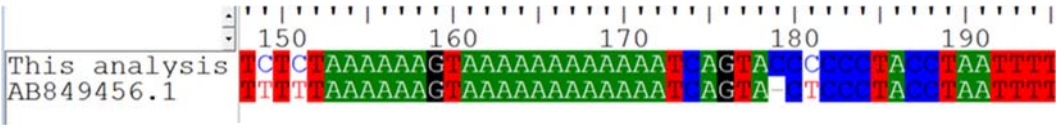

図 9. ゲンジと AB84945 の AT-rich region 配列をアライメント

ゲンジボタルでは、CDS 配列中の置換が数カ所見られた。本解析に用いた個体は奈良県、AB849456 は福井で採取されたものであるので、CDS 領域で見られた置換変異は地域差によるものだと考えられる。

次にヘイケボタルと AB84945 の比較を行った。ヘイケボタルは総塩基数が 16,719 bp であり、ゲンジよりも長くなっていた。これは、AT-rich region が長くなっていることが原因であった。ゲンジの AT-rich region が 1,369 bp であるのに対してヘイケでは 2,100 bp と、約 1,000 bp も長くなっていた。これは、近縁種の中でも二番目に長い長さとなっていた。

最後に近縁種との系統樹を作成した結果を示す（図 10）。ゲンジとヘイケは同じ *Luciola* 属であるが、分子系統樹上では属の異なる *Aquatica lei* の方が近縁であるという結果が得られた[2,3]。

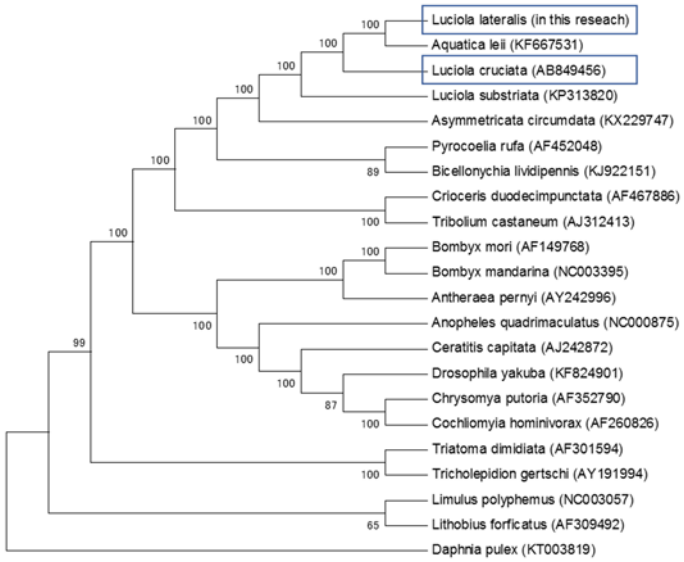

図 10. 近縁種との系統解析

4. まとめ

ヘテロ接合度の高さ、短いイントロン、GC 含量の低さ等、ホタルゲノムは他の生物に比べて非常に特徴的であることが分かった。全体の統合まではあともう一歩だが、ゲンジボタルは約 15,000 の遺伝子を持つことが分かった。

5. 謝辞

本研究課題を遂行するにあたり、研究助成を頂いた公益財団法人サンケイ科学振興財団に心から感謝申し上げます。

6. 引用文献

1. H. Amano et al., Complete mitochondrial genome sequence of *Luciola cruciata*. *Res. Rep. Fukushima Natl. Coll. Technol.*, **2013**, *54*, 149–152.

2. J. Maeda, et al., The complete mitogenome and phylogenetic analysis of Japanese firefly 'Genji Botaru' *Luciola cruciata* (Coleoptera: Lampyridae). *Mitochondrial DNA Part B.*, **2017**, *2*, 522-523.

3. J. Maeda et al., The complete mitochondrial genome sequence and phylogenetic analysis of *Luciola lateralis*, one of the most famous firefly in Japan (Coleoptera: Lampyridae). *Mitochondrial DNA Part B.*, **2017**, *2*, 546-547.

Genome analysis and gene comparison of the Japanese firefly, *Luciola cruciata*, for useful gene discovery

Dai-ichiro Kato

Department of Chemistry and Bioscience, Graduate School of Science and Engineering,
Kagoshima University
1-21-35, Korimoto, Kagoshima 890-0065 JAPAN
e-mail : kato@sci.kagoshima-u.ac.jp

Abstract

Approximately 40 types of fireflies are known in Japan. Among them, about half of the terrestrial and aquatic fireflies are discovered in Kagoshima prefecture and/or its island areas and each shows remarkable features. *Luciola cruciata*, one of the most famous fireflies in Japan, is known to be divided into six types depending on the difference of the mitochondrial COI gene. Although these show different emission cycles and oviposition manners, it has not be clear why these differences are caused. Thus to elucidate this mystery is necessary to analyze and compare the whole genome sequence. In this study, we attempted the genome analysis of *L. cruciata*. Research results indicated the distinctive properties different from other organisms in respects of high heterozygosity, short introns sequence, and low GC contents and so on. Genome annotation was performed by integrating three methods, RNAseq comparison method, Ab-initio method and homology search method. Although the integration was not completed within the research period, it could be predicted that *L. cruciata* has about 15,000 genes.