研究課題名：希少野生動物のゲノム解析

課題代表者：国立環境研究所生物・生態系環境研究センター　中嶋信美
共同研究者：国立環境研究所生物・生態系環境研究センター　大沼 学
酪農学園大学獣医学群　遠藤大二

実施年度：平成 25 年度～平成 25 年度

1. 研究目的

本年度、生物・生態系環境研究センターでは運用コストが低く高性能な次世代シーケンサーを導入した。さらに、当該シーケンサーの利用を動物園関係者に一部公開することにより、希少野生動物のゲノムデータが集積する基盤を整備した。ゲノム全長のデータは ENSEMBL サイトに 40 種の哺乳動物と 5 種の鳥類がすでに掲載されているが、野生動物の進化および種の保全に関する知識としての実用性を求めた場合、10,000 種が取得目標とされている（10k プロジェクト）。その実現のためには、次世代シーケンサーによる得られるランダムな塩基配列データを連結して基本的には重複の無い遺伝子情報として解析可能にし、さらに蛋白質機能についての相互比較を行う必用がある。そのための技術・プログラムは日々開発が進められ、オープンソースとして公開されているが、本研究では、次世代シーケンス、ゲノムおよび遺伝子分析プログラムをスーパーコンピュータ上に展開し、次世代シーケンサーデータを元に遺伝子の機能予測を可能にすることを目的とする。

2. 研究計画

最終的な機能解析を可能にするため、本年度においては次世代シーケンサーの生データを連結して遺伝子情報を含むゲノム断片とするアセンブラー、既知の塩基配列と比較するための相同性検索プログラムに加え、既知のゲノムデータに対し次世代シーケンサーデータを配置するためのアライメントソフトをスーパーコンピュータ上に配置し、一連の操作を行う体制を整備する。ゲノム分析については、多数のプログラムが公開され、ゲノム研究者がそれぞれ利用しているため、来年度においてプログラムの追加と連結した分析操作を可能にする。

3. 進捗状況

今年度においては、ヤンバルクイナゲノムから次世代シーケンサーより得られた 142GB 約 2.75 億本の塩基配列データを元にゲノムデータの構築を進めている。

従来の計算方法では、1,000GB 程度のメモリー容量を必用としたが、低メモリーで次世代シーケンサーからの生データを連結するプログラムとして SparseAssembler をスーパーコンピュータの前段階としてスカラー計算機の半分のメモリー量のパーソナルコンピュータで稼働した。結果として、1,770,973 本（平均 1,474bp）の連結配列（Contig）を構築した（図 1）。これらの Contig 配列を、鳥類の典型例としてのニワトリの遺伝子との相同性を確認したところ、「塩基配列が Contig 上に予測される遺伝子数の割合」は 88.4%となった。

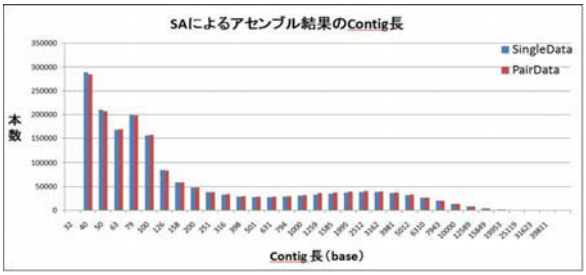

図 1　ヤンバルクイナゲノムデータのアセンブル結果。SparseAssembler により構築された Contig のサイズ分布。横軸にサイズ、縦軸にその本数を示す。アセンブル過程ではシーケンサーデータがある程度の長さのゲノム断片の両側で得られたことを仮定する PairData 解析と、そのような仮定を導入しない SingleData 解析が行われるが、本年度の解析では双方についてほぼ同様な結果が得られることが示された。

4. 今後の計画

アセンブル過程についてスカラー計算機で実際の計算を実施するとともに、ゲノムおよび遺伝子情報が公開されている鳥類の遺伝子との相同性検索を実施し、ヤンバルクイナの遺伝子群を、すでにゲノム全長が報告され遺伝子配列が推定されているニワトリ、七面鳥、キンカチョウ、ガチョウおよび ヒタキと比較し機能について推定する予定である。

5. 計算機資源の利用状況（2013 年 6 月～11 月）

実行ユーザ数: 3　CPU 時間 v_deb: 0 hours, v_cpu: 0 hours, v_8cpu: 0 hours, v_16cpu: 0 hours, 計: 0 hours