平成 23 年度 RICC 利用報告書

課題名（タイトル）：

全ゲノムシークエンスデータ解析パイプラインの スーパーコンピューター・理研 RICC 上での動作確認とチューニング

利用者氏名：○角田 達彦*, 藤本 明洋*, 阿部 哲雄*, 中村 英二**
理研での所属研究室名：
＊ 横浜研究所 ゲノム医科学研究センター 統計解析・技術開発グループ 情報解析研究チーム
****株式会社 ダイナコム 研究部**

１． 本課題の研究の背景、目的、関係するプロジェクトとの関係

近年の著しいシークエンス技術の発展により、個人ゲノムシークエンスが可能となった。超並列シークエンサーの発展は著しく、現在では 600Gbp（ヒトゲノムの約 200 倍）の塩基配列データが約２週間で得られている。全ゲノムシークエンスは世界中で活発に行われており、昨年は 1000 人ゲノム計画の論文が出版されたほか、がんの全ゲノムシークエンスも数多く報告されている。今後もシークエンサーのデータ産出量は増大していくことは確実であり、全ゲノムシークエンスは次世代の疾患研究において、極めて重要な役割を担っていくと考えられる。しかしながら、現在のシークエンサーには、読み取り長（リード長）が短い、エラー率が高いなどの問題があり、現在に至るまで解析手法が確立されているとはいえない。そこで、我々はシークエンサーからのデータの解析プログラムの開発を行った（Fujimoto et al Nat Genet 42: 931-936）。この解析パイプラインのさらなる高速化を行うため、RICC へ移植作業を行った。

２． 具体的な利用内容、計算方法

次世代シークエンサーから得られたリード配列を、標準ゲノム配列にマッピングし、確率計算に基づいて一塩基多様性の同定を行った。先行研究で解析したデータを用いて、RICC 上で解析を行った。

３． 結果

先行研究のデータの一部を用いて、解析パイプラインが正常に動作することを確認した。

４． まとめ

我々は、次世代シークエンサーの解析データを高精度に解析するパイプラインを構築し、RICC への移植作業を行った。小規模なテストデータを用いて、解析パイプラインが正常に動作することを確認した。

５． 今後の計画・展望

シークエンスコストの低下にともない、ゲノムデータの産出量は爆発的に増加している。今後は、シークエンスデータの大量解析に向けて、高並列化を行う必要がある。

６． RICC の継続利用を希望の場合は、これまで利用した状況（どの程度研究が進んだか、研究においてどこまで計算出来て、何が出来ていないか）や、継続して利用する際に行う具体的な内容

これまでの解析では、ヒトゲノム配列の一部のみを用いていたが、全ゲノムデータを用いた解析を行う必要がある。また、さらなる高速化に向けて、並列度を上げ、台数効果を高めるためのチューニングが必要であると考えられる。具体的には I/O を考慮した最適なデータ分割数の選定、異常終了する job の管理と自動再実行プログラムの開発が必要である。

７． 利用研究成果が無かった場合の理由

我々の解析パイプラインを用いた解析の成果は、２報の国際誌に掲載されている（Nat Genet 42: 931–936、Nature 464, 993-998.）。また、一報は現在審査中である。しかしながら、RICC の使用期間が短く、RICC を用いた研究成果は、まだありません。