HPCI 戦略プログラム分野 1「予測する生命科学・医療および創薬基盤」
課題 4「大規模生命データ解析」

スーパーコンピュータ「京」によるがん関連遺伝子の大規模高速検出システムの開発

Human Genome Center
the Institute of Medical Science, the University of Tokyo
THE INSTITUTE OF MEDICAL SCIENCE
THE UNIVERSITY OF TOKYO

HPCI 戦略プログラム 分野 1

概要

次世代シークエンサの急速な進歩に伴い，高速かつ安価にオミックスデータを得ることが出来るようになった（fig.1）．オミックスデータの解析（バイオインフォマティクス）には，アラインメント（リードの位置同定）や統計的検定などの計算処理が必要である．これらのビッグデータ処理を高速に実現するために，今日ではスーパーコンピュータは必要不可欠なインフラとなっている．
ほとんどの遺伝子解析ソフトウェアはパイプラインソフトウェア（外部ソフトウェアやスクリプト処理等を多段に組み合わせたソフトウェア）であり、逐次～数十並列程度の計算環境を念頭に置いた設計になっている．このため，「京」を始めとする大規模スーパーコンピュータ上での大規模解析は困難である．
本研究では,がん発生において重要な融合遺伝子検出パイプライン「Genomon-fusion」の「京」への移殖・高速化 (GFK:Genomon-fusion for K computer) に関する成果，および実際に大規模検体解析を実施した際の工夫などについて紹介する．

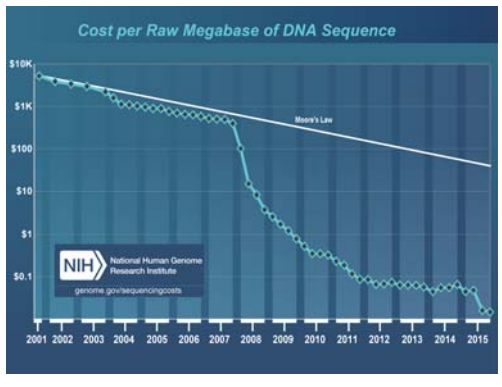

Fig.1 次世代シークエンサの進歩
1000 ドルゲノムを実現した Illumina 社の Hiseq X ten[*1] シリーズ（上）と 100 万塩基対読み取りに掛かるコストの年次変化[*2](左)．シークエンスコストは，次世代シークエンサの登場により急激に低下した．

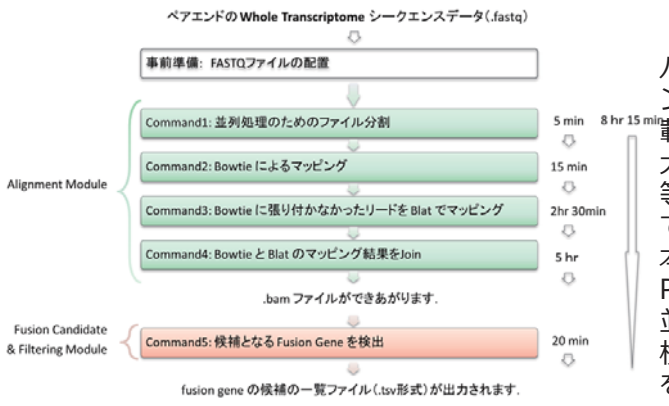

Fig.2 Genomon-fusion のワークフロー概要

パイプラインの再構成と並列化

バイオインフォマティクス分野で利用される計算機は，グリッドエンジン（小規模ジョブをまとめて並列に同時実行してくれるシステム）を搭載している．本機能は小規模解析を実施するユーザに便利である反面，大規模ジョブと混在した際にシステム稼働率の低下や待機時間の不平等等の問題を生じるため，「京」などのスーパーコンピュータには搭載されていない．
本研究では Genomon-fusion（fig.2) をアラインメント部，ソート・PCR 重複除去部，検出部の 3 パートに再構成し，各パートを MPI による並列プログラムとして実装した（fig.3）．並列化により，1 ジョブで数百検体の同時解析が可能となり，スーパーコンピュータ「京」のスケールを生かした大規模解析が可能となった．

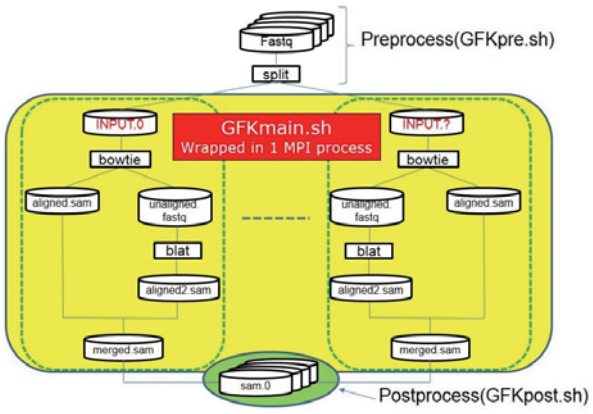

Fig.3 GFK のアラインメント部分概略図

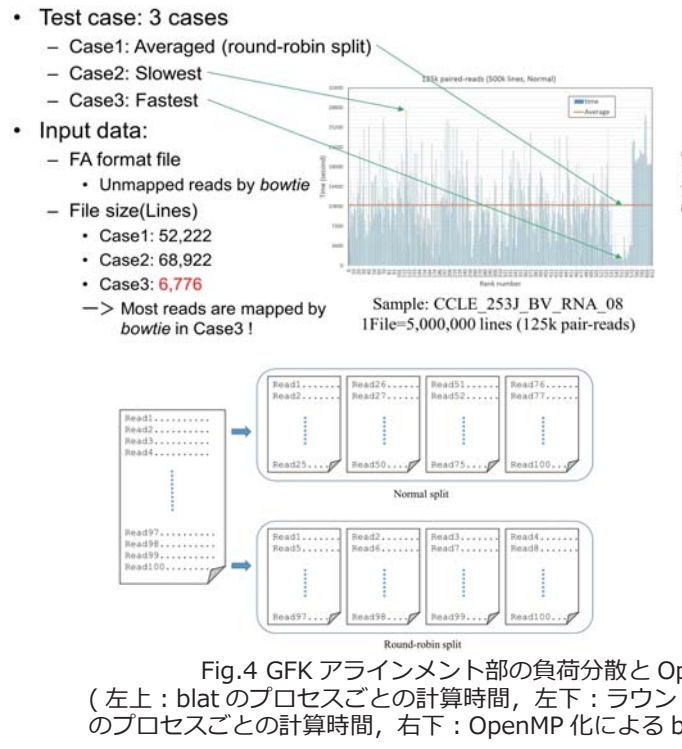

Fig.4 GFK アラインメント部の負荷分散と OpenMP 化による高性能化の結果
(左上：blat のプロセスごとの計算時間，左下：ラウンドロビン分割，右上：負荷分散後の blat のプロセスごとの計算時間，右下：OpenMP 化による blat のスピードアップの様子)

京の性能を生かすために：負荷分散と OpenMP 化

「京」の高い演算性能を引き出す (高性能化) ために重要な二つの技術が負荷分散と OpenMP 化である．本研究では，特に計算負荷の高いアラインメント部について行った改良について説明する．
まず初めに,パイプラインの各処理ごと･プロセスごとの計算時間を計測（プロファイリング）する．その結果，詳細マッピングに使用する blat が計算時間の大部分を占めており，プロセスごとの計算時間に非常に大きなばらつきがあることが分かった（fig.4 左上）．分析の結果，ごく一部のリード処理に多大な時間がかかっていることがわかったため，データ分割に工夫をする（fig.4 左下）ことにより，プロセス間での計算時間の均一化を実現した．
次に，blat 自体の高速化を検討した．blat は非常に多くのメモリを消費するソフトウェアであり,GFK では実行時に 8GB/process を割り当てている.「京」はノード当たり 8CPU コア・16GB の構成のため，ノード当たり 2 プロセスで計算を実行させていた．そのため，6 コアが余剰コアとして遊んでしまっていた．
OpenMP によるスレッド並列では，メモリ消費をあまり増やすことなく，余剰コアを有効活用することで処理時間の短縮が期待できる．いくつかのサンプルデータ（fig.4 左上）を用いた blat 内部のプロファイリングの結果，処理時間のほとんどを占めるループ処理を特定することに成功した．そのループ部分を OpenMP により自動スレッド並列化することで 1 プロセス当たり 4 コアでの計算が可能となり，約 2.5 倍（4 スレッド時）の高速化を実現した．
テストデータでの評価では，負荷分散と OpenMP 化の効果を組み合わせることにより 6.2 倍程度の高速化を達成した．

多数検体処理の意義とポスト「京」への期待

GFK による大規模検体解析を実施した．今回対象としたのは Cancer Cell Line Encyclopedia (CCLE) の RNA-seq データ 669 検体分である．この解析を,Shirokane2（東京大学医科学研究所ヒトゲノム解析センターのスーパーコンピュータ．2012 年 1 月～2015 年 12 月稼働．16,000 コア，約 150TFlops.）で解析した場合，システム全体を 3 か月程度占有するほどの解析規模である．「京」での解析では，のべ 160 万 CPU コア・93 万ノード時間のリソースを使用し，1 日ですべての計算を終了している．リソース消費量からの換算では,「京」フルノードの 24 時間占有時での解析能力は 11,422 検体となり，世界に類を見ない大規模検体解析システムとなっている[*3].
オミックスデータの解析は，今後ますます大規模化していくことが予想される．がんの遺伝子研究では，1% 程度の希少疾患がターゲットとされ，その解析には 1 万～ 2 万検体が必要と試算されている．また，米国の Precision Medicine に代表されるように，医療の個別化が今後の潮流であり，その中核技術の一つが遺伝子診断である．米国などではすでに DNA シークエンスから診断までを実施している病院もあり，オミックスデータの高速解析の必要性がすでに切実なものとなってきている．このような背景から，次期フラッグシップスーパーコンピュータであるポスト「京」の膨大な計算資源を活用した大規模かつ高速なオミックスデータ解析システムへの期待は非常に大きい．
本研究はポスト「京」における重点課題 2「個別化･予防医療を支援する統合計算生命科学（課題責任者：宮野悟)」と密接に連携しており，戦略分野での知見・成果等をコデザインに反映している．

Tab.1 CCLE 検体解析の概要

	Elapse time (hours)	CPU core
GFKalign	875,122.2	1,574,540
GFKdedup	4,512.8	669
GFKdetect	53,143.0	29,436
Total	932,778.0	1,604,645

参考文献
1. http://jp.illumina.com/systems/hiseq-x-sequencing-system/system.html
2. http://www.genome.gov/sequencingcosts/
3. Ito, S. et al., 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp.1441-1447, 2015.

宮野　悟（課題代表），伊東　聰（研究担当者）
東京大学医科学研究所　ヒトゲノム解析センター
E-mail: miyano@ims.u-tokyo.ac.jp
sito@hgc.jp

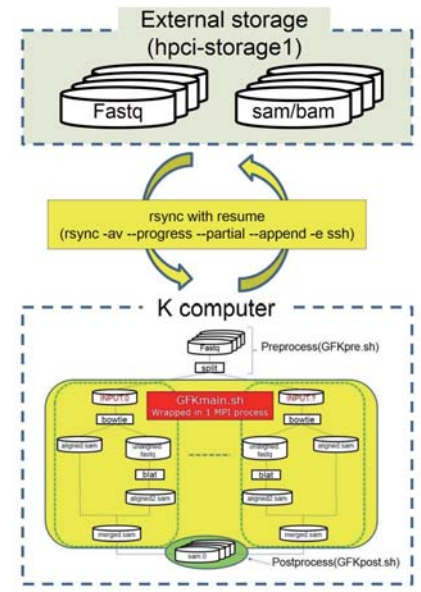

Fig.5「京」と外部ストレージとの協調