Vol.2019-HPC-171 No.8
2019/9/20

バイオインフォマティクス分野における HPC 的取り組みの紹介

伊東　聰[1,a)]　矢留 雅亮[1]　宮野 悟[1,b)]

概要：バイオインフォマティクスは，ビッグデータを解析する分野の一つとして広く知られている．近年，次世代シーケンサー（Next Generation Sequencer:NGS）等の機器の飛躍的な性能向上により，本分野で取り扱うデータ量は膨大なものとなってきた．これに伴い，旧来の PC による逐次的な解析では処理が間に合わず，スーパーコンピュータ，GPGPU，FPGA 等を利用した HPC への取り組みが見られるようになってきている．本稿では，著者らが実施した研究を元に本分野を HPC 研究者に紹介し，新規参入を促すことを目的とする．

Introduction of HPC studies in Bioinformatics

1. はじめに

バイオインフォマティクスとは，DNA を始めとする生体情報を数理的に解析する学問分野である．21 世紀に入り，実験機器の性能が劇的に向上した [1] おかげで，本分野は飛躍的に発展している．とりわけ，NGS の出力するオミックスデータを利用した研究の進歩は著しい．**図 1** は米国 National Institute of Health が公開しているシーケンスコストの推移を示すグラフである．1990 年に開始された米国のヒトゲノム計画 [2] は 13 年もの年月と 30 億ドルに及ぶ膨大な資金とを要するものであった．2019 年現在，最新のシーケンサーでは同じヒトゲノムの読み取りが 1000 ドル以下，46 時間程度で可能であることと比較すれば，どれほど進歩がめざましいものであるかが理解できるだろう．機器の進歩に伴い生じてきた新たな問題が，データ処理に要する計算機の性能である．

バイオインフォマティクスでは，実験で得られたデータを統計的に処理する必要がある．これまで取り扱ってきたデータは，サンプル数的にもデータサイズ的にも比較的小さく，PC での解析で十分に対応可能なレベルであった．しかし，2019 年現在，Nature や Science 等のトップジャーナルに掲載される最先端研究では，1 サンプル当たりのデータ量はテラバイト前後まで増大し，対象とするサンプル数は大きなものでは 1 万を超えるようになっている．もはや PC で扱えるデータ量を超えていることは自明であり，AWS などのクラウドリソースやスーパーコンピュータの利用が必須である．

しかしながら、クラウドリソースに比べてスーパーコンピュータを利用したバイオインフォマティクス研究例はそう多くない．その原因は，スーパーコンピュータの運用形態とバイオインフォマティクスに携わる研究者の性質の二つの側面が考えられる．本稿では，著者らが主に取り組んできたがんのオミックスデータ解析パイプラインのスーパーコンピュータ「京」への移殖と高速化などを通して得

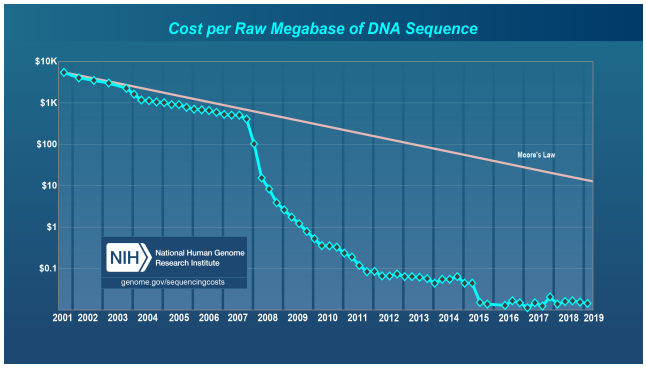

図 1　Mb 当たりのシーケンスコストの推移 [1]
Fig. 1　Sequencing cost per Mb[1]

[1] 東京大学医科学研究所
The Institute of Medical Science, The University of Tokyo, Tokyo 108–8639, Japan
[a)] sito@hgc.jp
[b)] miyano@ims.u-tokyo.ac.jp

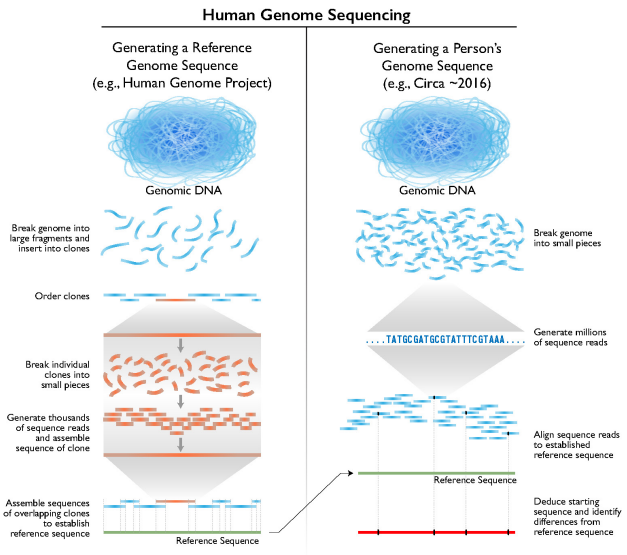

図 2 リファレンスゲノムとアラインメント [1]
Fig. 2 Human genome sequencing[1]

られた知見をもとに，既存のスーパーコンピュータ利用分野のソフトウェアとの違いを説明する。

2. オッミックスデータ解析に関する基礎知識

2.1 用語定義および説明

bp base-pair（塩基対）の略.

リード 断片化された DNA や RNA のうち，シーケンサーに読み取られる部分.

デプス DNA や RNA のある塩基対の位置におけるリードの本数.

PCR Polymerase Chain Reaction の略．対象の DNA などを増幅する.

がん遺伝子 DNA に変異が入ることにより正常細胞をがん化させる要因となった遺伝子.

germline/somatic それぞれ生殖細胞系列および体細胞系列.

2.2 解析前半：アラインメント

オミックスデータ解析は主として前半と後半に分けることが出来る．前半部分はアラインメント処理が主であり，後半部分が各種変異検出処理となる.

アラインメント（マッピングとも言う）とは，シーケンサーが読み取った塩基配列情報を標準 DNA 配列情報等をリファレンスとして，その位置情報を確定する作業（**図 2**）を意味する.

現在主流であるシーケンサーは sequence by synthesis 法（SBS）[3] を用いている．本手法は読み取る塩基配列に蛍光分子を結合させることで光学的に塩基を特定する．この処理は酵素を用いた生化学反応であり，1 塩基の読み取りに蛍光分子の着脱とレーザー励起発光の光学撮影の時間を

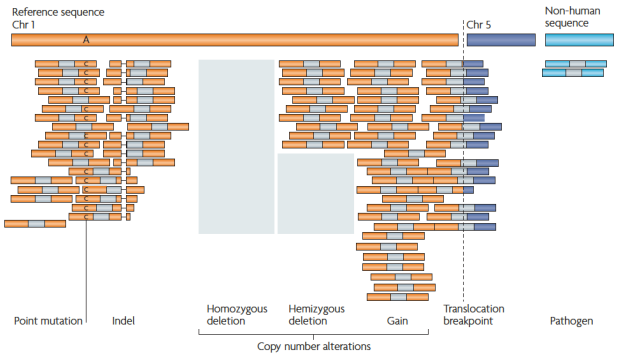

図 3 代表的な変異の例 [4]
Fig. 3 Major variations in human DNA[4]

必要とする．処理時間は塩基長に比例して長くなってしまうため，処理の高速化を目的として長い塩基配列を断片化（Fragmentation）している．シーケンサーの読み取る塩基配列情報は，この断片化された DNA 等の片側または両側の 100 - 200 塩基程度であり，この部分をリードと呼ぶ．その結果，位置情報の特定されたリードの集合情報から，元の塩基配列情報を再構成することが必要になるのである.

アラインメント後のリードは，後半の変異検出処理のために事前処理を施す．必須処理としてはソートと PCR 重複除去である．ソートは文字通り，アラインメント情報に基づいてリードを並べ替えるだけである．PCR 重複除去では，同一判定されたリードを除去する．本処理は，後半の変位検出の精度に影響する重要な処理である．PCR による断片の増幅は，シーケンサーによるリードの捕捉率や SBS における発光強度を高める目的で実施されるが，時折，異常増幅される断片が存在する．この偏りは，変異の存在確率の計算結果に多大な影響を与えることは自明なため，同一リードを除去することによってこの偏りを是正している．これらの処理を施したリードは，最終的に SAM またはその圧縮形式である BAM フォーマットとして 1 ファイルに集約される.

2.3 解析後半：変異検出処理

本処理は対象とする変異によって処理内容が変わる．**図 3** に代表的な DNA の変異の例を示す．特定の部位のみを観察することで変異を把握できるもの（ex. Point mutations：点突然変異）もあれば，2 つの染色体にまたがるような変異（ex. Translocation：転座）もあることが分かる．各変異は，SAM/BAM ファイルからその特徴を含むリードを抽出することで検出することが出来る．ここで重要な点は，検出した変異の偽陽性（False Positive:FP）確率である.

点突然変異を例に説明する．点突然変異とは，ある 1 部位の塩基が別の塩基に変わる変異である (図 3 では，対象部位の塩基が C に変化している).

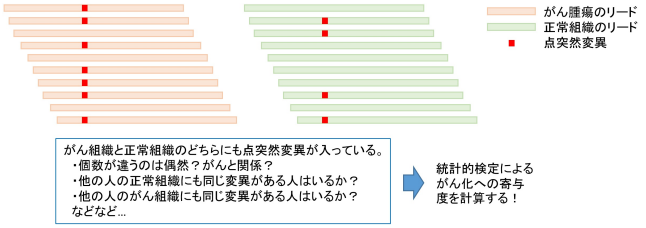

図 4 シーケンス結果における点突然変異の見え方の例 [5]
Fig. 4 An example of point mutations in sequenced results[5]

表 1 図 4 における変異分布
Table 1 Statistics of 図 4.

	Tumor	Normal	Total
mutation reads	7	4	11
normal reads	3	6	9
Total	10	10	20

がんを念頭に置くと，この変異ががんの発生や進化に寄与している場合にこの変異は重要となる．DNA は個々人で異なるものであるから，一般化されたリファレンスの DNA 配列情報と違うだけでは，それが患者の個性かがん遺伝子の原因かが区別できない．また，DNA は同じ固体の細胞間でも微妙に異なる（不均一性）．さらに，シーケンサーの読み取りエラーも低確率ではあるが含まれる．候補変異が FP かどうかは統計的手法（Fisher 検定など）を用いて確率的に判定し，最終的には実験的に確認することになる．

図 4 の例で具体的に Fisher 検定をしてみよう．Fisher 検定では，検定したい項目（ここでは点突然変異候補）が二つの集団（正常細胞とがん細胞）で統計的に差があるかどうかを検定する．図 4 の例では，変異の有無によるリード本数の分布は**表 1** のようになっている．

点突然変異の検討塩基位置に存在するリードの本数をデプスという．本例の場合はデプス 10 である．この分布に対し Fisher 検定を実施してみると，およそ 0.3698 という数値が得られる．この値（p 値）が低ければ低いほど，二つの集団における検討項目の数値の差が偶然には発生しないことを意味し，真陽性（True Positive:TP）であると判断できる．では，割合が同じでデプスが 10 倍，すなわちリード数がそれぞれ 100 本ずつシーケンスされていた場合の p 値はというと，3.304×10^{-5} と桁違いに小さくなる．つまり，利用できるデータ量が多ければ多いほど指数関数的に検定の精度を上げることが出来る．このため，本分野では検体（サンプル）数はもちろん，1 検体のデータ量も年々増加している．

図 5 に Stephens ら [6] がまとめた，全世界での DNA シーケンスのデータ量統計と今後の予測量のグラフを示す．現在の推移のまま増加すれば 2025 年には 1Zbp を超えてしまう．実際，2015 年に発表された米国オバマ大統

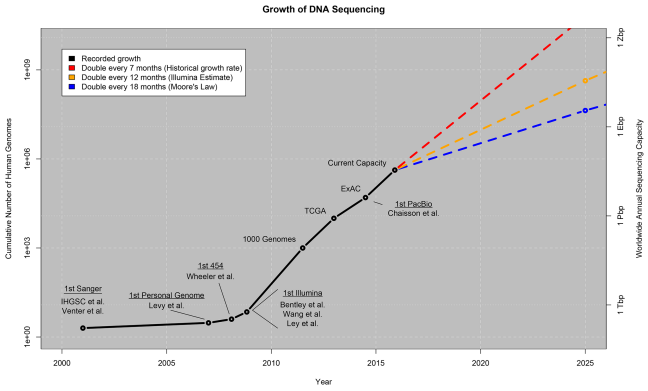

図 5 DNA シーケンス量の推移と予測 [6]
Fig. 5 Growth of DNA sequencing.[6]

領（当時）の Precision Medicine Initiative に基づき，各国で医療現場における全ゲノムシーケンスの利用（Clinical Sequencing）が活発化しており，シーケンス量は爆発的に増えている．

3. オミックスデータ解析の HPC

3.1 既存パイプラインの HPC 化状況

オミックスデータ用ソフトウェアは 2 種類に大別され，一つはアラインメントやデータハンドリングなどの特定用途用ソフトウェア，もう一つはターゲット変異を抽出する統合解析パイプラインである．前者の例としては BWA（アラインメント）[7]，bowtie2（アラインメント）[8]，samtools（SAM/BAM ファイルハンドリング）[9]，Picard（SAM/BAM ファイルハンドリング）[10]，後者には GATK（germline/somatic 変異検出），MuTect（somatic 変異検出）[11] などがある．

特定用途用ソフトウェアは開発者が一人ということも珍しくなく，ハイレベルなチューニングが施されていることもある（BWA など）一方で，スレッド並列化等も未実装なソフトウェアもあり，HPC 的な観点からはソフトウェアごとの性能ばらつきが非常に大きい．

統合解析パイプラインでは，グリッドエンジン（Univa Grid Engine[12] など）を前提とした分散コンピューティングの実装は一般的であるが，パイプライン全体のパフォーマンスについて十分な検討がなされているものはごく僅かである．

高度な HPC 化の例としては，NVIDIA の実施した GPU 移植版 bowtie2 である nvBowtie[13]（CPU の 3 倍前後）がある．また，GATK については Parabrics 社の GPU 版 [14] と Illumina 社の FPGA 版である DRAGEN[15] がある．

3.2 Genomon の HPC 化

本節では我々が実際にパイプラインの高速化を実施した例を紹介する．Genomon は東京大学医科学研究所で開発されている変異解析パイプライン群である．初期バージョ

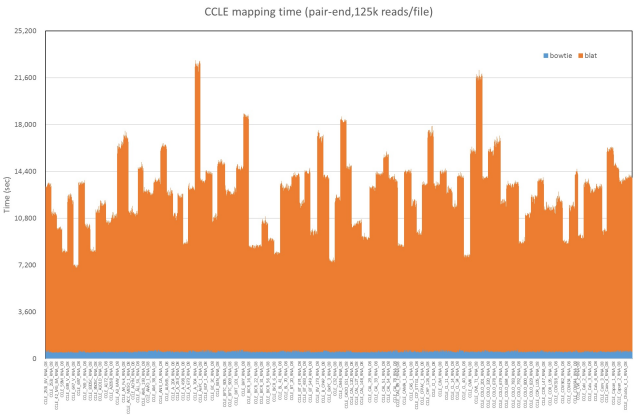

図 6 100 サンプルのアラインメント時間の内訳
Fig. 6 Breakdown of alignment time for 100 samples.

```
static struct dnaSeq seq;
struct
lineFile *lf = lineFileOpen(fileName,TRUE);

while( faMixedSpeedReadNext(lf,&seq.dna,
                    &seq.size,&seq.name) )
{
    searchOneMaskTrim( &seq, isProt, gf,
      outFile,maskHash, &totalSize, &count );
}

lineFileClose(&lf);
```

図 7 blat のホットスポット部分のソースコード
Fig. 7 Source code of blat hot spot.

ンでは exome 用，RNA-seq 用など，解析用途ごとに独立したソフトウェアとして公開されていたが，最新バージョンでは統合解析パイプラインとして全機能がまとめられている．詳細についてはプロジェクトウェブサイト [16] を参照されたい．

3.2.1 ホットスポットソフトウェアの単体チューニング

1 例目は 2015 年当時に利用されていた統合前のパイプラインの一つである Genomon-fusion のチューニングに関するものである．本内容の詳細は既報（[17], [18]）を参照されたい．本パイプラインは融合遺伝子を検出する．特徴として，検出の高速化と高精度化を両立するためにアラインメント部分に bowtie と blat[19] の二つのソフトウェアを利用している点がある．高速かつ広範囲をカバーする bowtie で大部分の正常なリードを処理し，融合遺伝子を含む可能性の高い残りの僅かなリードに対してのみ，blat を用いて正確なアラインメントを実施している．

Genomon-fusion も他のパイプライン同様，アラインメント部分に使用する計算機リソースの大部分が集中している．なかでも blat の占める割合が非常に高く（**図 6**），本部分の高速化がパイプライン全体の効率を左右するホットスポットとなっている．

アラインメントは本質的にメモリ消費量が多くなる傾向がある．なぜなら，インプットデータ以外にリファレンスデータが必要なためであり，インプットデータが 100MB/プロセス程度なのに対し，リファレンスデータは 5GB〜10GB 程度にもなってしまうからである．リファレンスデータは理論上，分割することは出来ず固定値である．このため，スレッド数を増やすことでスレッド当たりのメモリ消費量を下げることは重要である．

アラインメントは演算がリードごとに独立しており，理論的には並列化に向いている．そのため，bwa のようにスレッド並列化に対応したソフトウェアもあるが，本ソフトウェアはスレッド並列化は未実装であった．blat のソースコードに対して区間ごとの実行時間を計測したところ，幸

```
#pragma omp parallel private(i,ii,thread_num)
{
    thread_num = omp_get_thread_num();

#pragma omp for  // Mail loop
    for( ii = 0; ii < lcount; ii++ ) {

    searchOneMaskTrim(&seq[ii], isProt, gf,
    outFile[thread_num], maskHash, &totalSize,
     &count, thread_num);
    } // End of main loop
} // End of OpenMP region
```

図 8 OpenMP 化後のホットスポット部分のソースコード
Fig. 8 OpenMP parallerized hot spot source code of blat.

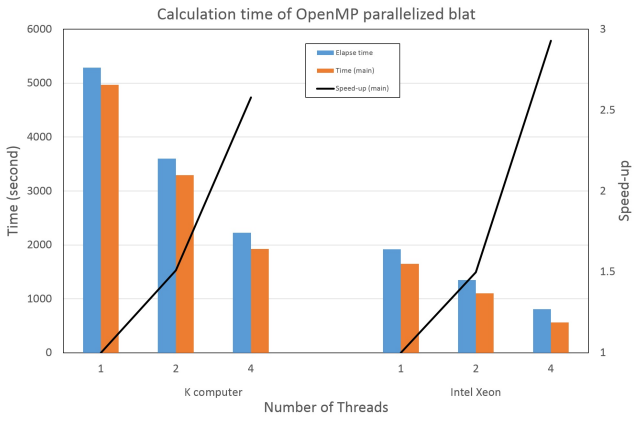

図 9 blat のスレッド並列による計算時間変化
Fig. 9 Results of OpenMP thread parallelization of blat.

いなことに 90%前後を占めるホットスポットループを発見（**図 7**）した．京コンピュータ 1 ノードおよび Intel マシン（E5450 3.0GHz）でテストしたところ，OpenMP によるスレッド並列化（**図 8**）により 4 スレッド程度まではスレッド並列の効果が観測できた（**図 9**）．本例のように，ごく一般的な時間測定とスレッド並列化だけでも効果的な高速化が出来る場合がある．

3.2.2 パイプライン全体のロードバランシング

前項では独立したソフトウェアに対するチューニング例を示したが，ここではパイプライン全体の特性とその高速化に関する取り組みについて紹介する．本項の例も詳細に関しては既報（[17], [18]）を参照されたい．

バイオインフォマティクス分野のパイプラインはグリッドエンジンを前提とした分散コンピューティングに対応していることについては前述のとおりである．しかし，その並列化はあくまで 1 検体の計算に対してのみ有効であり，多数検体の計算全体に関しては個別にジョブ投入することを前提とした設計である．これはグリッドエンジンの特性に沿ったソフトウェア形態の帰結でもある．

一方，大規模スーパーコンピュータは小さなジョブを大量に投入する用途には不向きであり，MPI を用いて多数検体を同時解析するようにパイプラインを修正する必要がある．

京コンピュータを用いた大量検体解析を目的とした Genomon-fusion の京コンピュータへの移植（以降，Genomon-fusion for K computer:GFK）に際し，我々はパイプラインを 3 つ（順に GFKalign，GFKdedup，GFKdetect）に再構成し，それぞれを MPI により並列化した．

この並列化・高速化の目的は，多数検体の解析を完了するまでの時間を短くすることである．この観点での最大の課題は GFKdedup であった．本処理は 1 検体に対して 1 プロセスの逐次処理であるが，1 検体に対するパイプライン全体の経過時間のうち 70%程度を占めていた．本処理は，GFKalign にて並列にアラインメントされた全リードデータに対しソートや PCR 重複リードの除去等を実施するプロセスであり，アルゴリズム的には並列化することは不可能である．

そこで我々は処理内容の生物学的意味を検討した．その結果，染色体レベルでのデータの分割が理論的に可能であることを確認し，逐次処理を 26 並列（人間の染色体対数）へと並列化することに成功した．その他のチューニングと合わせた結果，最終的に京コンピュータ全ノード利用時には 5,356 検体/日の大規模解析能力を達成した．

4. バイオインフォマティクス分野の HPC 促進に向けて

本分野の発展は著しく，スーパーコンピュータと HPC の需要はこれから爆発的に増えてくることは明白である．前節にて HPC 化の例を示したが，今後新たに参入する HPC 研究者に向けて，著者らの知見を紹介して本稿のまとめとしたい．

4.1 非同期性とロードインバランス

HPC 的観点から本分野のソフトウェアの特徴を上げるならば，非同期性と均一化の困難なロードインバランスの

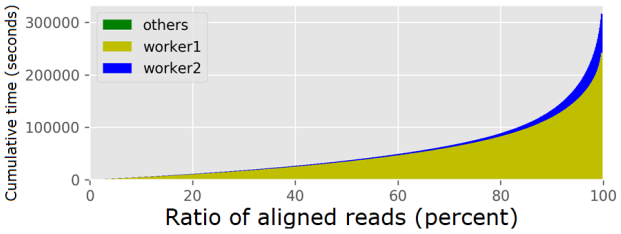

図 10 ATL 検体に対する BWA によるアラインメント積算時間の推移

Fig. 10 Cumulative time of BWA alignment for an ATL sample.

2 点である．大規模並列計算が普及している物理，工学分野のシミュレーションにおいて計算量の目安となるのは格子点や要素などの自由度（DOF）である．計算量は次数や要素により異なるものの，自由度あたりの計算量は概ね一定である．よって，並列計算する際の計算負荷は領域分割法などで 1 プロセスあたりの自由度数を均一化することにより実現可能である．また，一つの自由度における計算に他の自由度における物理量が必要なことから，並列計算の際に通信によるプロセス間でのデータ補間が必要になる．

一方，本分野における解析で自由度に相当するものはリード，すなわち DNA や RNA の断片である．リードに対する演算は基本的に独立しているため，データ分割に起因するプロセス間の通信は基本的には不要なことから Embarassingly parallel（EP）である．ところが，1 自由度，すなわちリードに対する計算負荷は非常にばらつきが大きい．例えば，著者らが成人 T 細胞白血病リンパ腫の検体 [20] で BWA によるリードごとの処理時間を調べたところ，検体によらず全体の約 5%程度のリードが全処理時間の 40%程度を占めるという結果であった（**図 10**）．

EP であるがゆえに並列化が容易である一方，著しい負荷の非均一性が実際には並列化による高速化を阻害しており，高い並列効率の実現は非常に難しいことが分かる．

また，負荷の非均一性はパイプラインを構成する複数のタスク間にも存在する．パイプラインを構成する各タスクは逐次から数千並列まで，処理内容によって並列度が大きく変化する．また，CPU コアリソース消費量は大きいものの並列化により経過時間は大きくないタスクもあれば，逐次処理だが経過時間が非常に大きいタスクもある．

パイプラインの HPC 化において，MPI によるパイプラインの並列化が第一選択肢となる可能性が高いが，タスク間の計算負荷非均一性を十分に考慮し，パイプラインの分割等を適切に実施する必要がある．場合によっては，プログラムのアルゴリズムだけでなく，生物学的な知識も合わせた改良を考慮しなければならない．

4.2 ファイルシステムへの負荷

パイプライン中のタスク間におけるデータ授受の基本はファイル渡しである．クラウドシステムやグリッドエンジン環境下のスパコンでは，並列プロセスの開始時刻が資源の空き状況により同期しないため，Lustre 等の分散ファイルシステムのボトルネックである MDS に負荷がかかることは稀である．一方，パイプラインを MPI 並列化し，複数検体解析を 1 ジョブに包含するような運用をした場合，MPI の同期特性から並列度の高いタスクの開始および終了時に一斉にファイルアクセスが生じ，MDS の処理可能要求数を超える可能性に注意しなければならない．

4.3 その他

HPC 研究者が本分野に参入する場合は医学・生物学者との共同研究になることが前提と考えられる．一般的な傾向として，医学・生物学者は計算機利用に関する高度な知識を取得する意欲は少ない傾向にある．バイオインフォマティクスと呼ばれる本分野の研究には，医学，生物学，化学等の基礎学問に加え，患者や動物実験に関連する倫理法規制への理解，さらに，得られた実験データの解釈のために統計学までカバーする必要があるためである．スーパーコンピュータの普及の観点からは，並列化やチューニングの手法を単に提示するだけでは不十分であり，ライブラリやパッケージ等の完全な利用環境の状態まで完成させた上で公開・提供することが望ましい．

5. おわりに

本稿では，バイオインフォマティクス分野における HPC 研究の事例として，著者らが行ったパイプライン Genomon の京コンピュータへの移植研究からソフトウェアの単体チューニングおよびパイプラインのロードバランシングの内容を紹介した．バイオインフォマティクス用パイプラインはフローが複雑なことが多いため，HPC 化の際の参考にしていただければ幸いである．

また，グリッドエンジンの無いスーパーコンピュータ上へパイプラインを移植し，運用する際の問題点についても述べた．

グリッドエンジンを用いるジョブと MPI も利用するジョブは資源の利用方法が違うため，共存が難しい．著者らはパイプラインの移植・運用を容易にする仮想的なグリッドエンジン環境 Virtual Grid Engine（[21], [22]）を開発したが，将来的には MPI ジョブで消費しきれない遊休コアをグリッドエンジンジョブが利用するような，両タイプのジョブが共存するスケジューラが開発されることを期待したい．

謝辞 本研究は，ポスト「京」（スーパーコンピュータ「富岳」）で重点的に取り組むべき社会的・科学的課題に関するアプリケーション開発・研究開発における重点課題 2「個別化・予防医療を支援する統合計算生命科学（課題番号：hp190158）（課題責任者：東京大学・宮野悟）の支援を受け実施したものである．

参考文献

[1] Wetterstrand, K. A.: The Cost of Sequencing a Human Genome, National Human Genome Research Institute (online), available from ⟨https://www.genome.gov/sequencingcosts/⟩ (accessed 2019-07-10).
[2] International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome, *Nature*, Vol. 409, No. 6822, pp. 860–921 (online), DOI: 10.1038/35057062 (2001).
[3] Canard, B. and Sarfati, R. S.: DNA polymerase fluorescent substrates with reversible 3′-tags, *Gene*, Vol. 148, No. 1, pp. 1 – 6（オンライン）, DOI: https://doi.org/10.1016/0378-1119(94)90226-7 (1994).
[4] Meyerson, M., Gabriel, S. and Getz, G.: Meyerson M, Gabriel S, Getz GAdvances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11(10): 685-696, *Nature reviews. Genetics*, Vol. 11, pp. 685–96 (online), DOI: 10.1038/nrg2841 (2010).
[5] 伊東 聰，矢留雅亮：がんの遺伝子解析とスーパーコンピューティング，ポスト「京」重点課題 2 ニュースレター 6，東京大学 医科学研究所，東京都 (2017).
[6] Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S. and Robinson, G. E.: Big Data: Astronomical or Genomical?, *PLOS Biology*, Vol. 13, No. 7, pp. 1–11 (online), DOI: 10.1371/journal.pbio.1002195 (2015).
[7] Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *arXiv e-prints*, p. arXiv:1303.3997 (2013).
[8] Langmead, B., Wilks, C., Antonescu, V. and Charles, R.: Scaling read aligners to hundreds of threads on general-purpose processors, *Bioinformatics*, Vol. 35, No. 3, pp. 421–432 (online), DOI: 10.1093/bioinformatics/bty648 (2018).
[9] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, . G. P. D. P.: The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, Vol. 25, No. 16, pp. 2078–2079 (online), DOI: 10.1093/bioinformatics/btp352 (2009).
[10] Broad Institute: Picard toolkit, Broad Institute (online), available from ⟨http://broadinstitute.github.io/picard/⟩ (accessed 2019-07-26).
[11] Cibulskis, K., Lawrence, M., L Carter, S., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., S Lander, E. and Getz, G.: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples, *Nature biotechnology*, Vol. 31, pp. 213–219 (online), DOI: 10.1038/nbt.2514 (2013).
[12] Univa Corporation: Univa Grid Engine, Univa Corporation (online), available from ⟨http://www.univa.com/products/⟩ (accessed 2019-07-26).
[13] NVIDIA: NVBIO, NVIDIA (online), available from ⟨https://developer.nvidia.com/nvbio⟩ (accessed 2019-07-26).
[14] Parabrics: Parabricks Germline Pipeline, Parabrics (online), available from ⟨https://www.parabricks.com/germline/⟩ (accessed

2019-07-26).

[15] Miller, N. A., Farrow, E. G., Gibson, M., Willig, L. K., Twist, G., Yoo, B., Marrs, T., Corder, S., Krivohlavek, L., Walter, A., Petrikin, J. E., Saunders, C. J., Thiffault, I., Soden, S. E., Smith, L. D., Dinwiddie, D. L., Herd, S., Cakici, J. A., Catreux, S., Ruehle, M. and Kingsmore, S. F.: A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases, *Genome Medicine*, Vol. 7, No. 1, p. 100 (online), DOI: 10.1186/s13073-015-0221-8 (2015).

[16] Chiba, K., Okada, A. and Shiraishi, Y.: Genomon - The Zen of Cancer Genome Sequence Analysis, Team Genomon (online), available from ⟨https://genomon-project.github.io/GenomonPagesR/⟩ (accessed 2019-07-26).

[17] Ito, S., Shiraishi, Y., Shimamura, T., Chiba, K. and Miyano, S.: High performance computing of a fusion gene detection pipeline on the K computer, *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1441–1447 (online), DOI: 10.1109/BIBM.2015.7359888 (2015).

[18] 惣一朗鈴木，聰　伊東，奈生池田，Balazs, G.，郁夫三吉，直也丸山，真一朗滝澤，洋介村瀬，裕　石川，悟　宮野：ヒトゲノム解析プログラム Genomon-fusion の大規模スーパーコンピュータ向け最適化と性能モデル化，ハイパフォーマンスコンピューティングと計算科学シンポジウム論文集，Vol. 2016, pp. 15–26 (2016).

[19] Kent, W. J.: BLAT—The BLAST-Like Alignment Tool, *Genome Research*, Vol. 12, No. 4, pp. 656–664（オンライン），DOI: 10.1101/gr.229202 (2002).

[20] Kataoka, K., Nagata, Y., Kitanaka, A., Shiraishi, Y., Shimamura, T., Yasunaga, J.-I., Totoki, Y., Chiba, K., Sato-Otsubo, A., Yamamoto, S., Ishii, R., Muto, S., Kotani, S., Watatani, Y., Takeda, J., Sanada, M., Tanaka, H., Suzuki, H., Sato, Y., Shiozawa, Y., Yoshizato, T., Yoshida, K., Makishima, H., Iwanaga, M., Ma, G., Nosaka, K., Hishizawa, M., Itonaga, H., Imaizumi, Y., Munakata, W., Ogasawara, H., Sato, T., Sasai, K., Muramoto, K., Penova, M., Kawaguchi, T., Nakamura, H., Hama, N., Shide, K., Kubuki, Y., Hidaka, T., Kameda, T., Nakamaki, T., Ishiyama, K., Miyawaki, S., Yoon, S. S., Tobinai, K., Miyazaki, Y., Takaori-Kondo, A., Matsuda, F., Takeuchi, K., Nureki, O., Aburatani, H., Watanabe, T., Shibata, T., Matsuoka, M., Miyano, S., Shimoda, K. and Ogawa, S.: Integrated molecular analysis of adult T cell leukemia/lymphoma, *Nature Genetics*, Vol. 47, pp. 1304–1315 (2015).

[21] Ito, S., Yadome, M., Nishiki, T., Ishiduki, S., Inoue, H., Yamaguchi, R. and Miyano, S.: Virtual Grid Engine: Accelerating thousands of omics sample analyses using large-scale supercomputers, *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 387–392 (online), DOI: 10.1109/BIBM.2018.8621285 (2018).

[22] Ito, S.: Virtual Grid Engine, Human Genome Center, The Institute of Medical Science, The Univ of Tokyo (online), available from ⟨https://github.com/SatoshiITO/VGE⟩ (accessed 2019-07-26).