講演番号 10-DA02

大規模生物データ処理のための並列データベース

森下真一（東京大学）

概要　近年， DNA解読装置の革命的進歩（ゲノム情報ビッグバン）により，ムーアの法則を凌ぐ約8カ月で倍の速度で解読速度が上昇している. 2003年に解読が宣言されたヒトゲノムには約500万ドルの費用がかかったが，2010年現在では数日で数万ドル程度，2-3年後には1時間以内に1000ドル以下で解読が可能になると言われている.

DNAデータ処理における基本的なソフトウェアに，DNAアセンブラとアラインメントがある. いまだに研究の余地があるものの, おおよそ解決の見通しはついてきている. いま最も困難な問題となっているのは巨大ゲノムデータの高速処理である. 2010年度末には1日当たり250億塩基の解読能力のある装置が普及しつつある. ヒトゲノムの場合，両親に由来する2つの染色体の差を識別するために，1人から約1000億塩基を収集する．塩基情報に加えて，塩基の読取信頼度情報が付加されるので，1塩基あたり1バイトの記憶領域が必要とすると 100GB/人の情報が蓄えられる. このデータは多数のDNA断片情報であり，ディスクから読み込み，ヒトゲノム上でアラインメントし，位置に関する情報を再びディスクへと書き込むが，書き込みデータ量は入力データ以上に大きくなる．多数の CPU から大規模データの読み込みと書き込みが並列して実行されると，CPUとディスク間のデータ転送量がボトルネックとなる. ブロックサイズをある程度大きくし，ランダムアクセスは避け，シーケンシャルなデータ転送による実装を心がけると，200～500MB/秒程度の転送量を確保できそうである.

現在の超並列計算機システムでは，どうしても主記憶，CPU，主記憶間のデータ転送の性能向上に優先度が置かれていて，主記憶とディスク間のバッファ管理に配慮したシステムが少ない．そのため，DNA 情報処理には不向きなシステムが多い．2013～4 年ごろには 1 日当たり 1TB の情報量を生む解読装置が普及し，数万人分のヒトゲノム情報を処理する時代がくる可能性もある. ファイルアクセスが快適な並列計算機の構築が，DNA分析の鍵になる．本研究ではこれらの問題の解決に取組む.

1. 研究の目的と意義

DNA解読の応用は，以下に示すように生命科学全体にわたって多岐にわたる.

家族性癌など遺伝の要素の強いさまざまな病気や，薬の効果の個人差，酒の上戸・下戸（アルコール分解能力の個人差), 集中力の個人差などはゲノムの個体差に由来すると考えられている．こうした個体差は様々であり，遺伝子の塩基配列のうち，一つの塩基が別の塩基に置換されていたり，一部の塩基配列が欠落／挿入されたり，ゲノム中にコードされる遺伝子のコピー数の違い，短い塩基配列の重複数の違いといったことから生まれる.

最初に解析されたヒトゲノムは，不特定多数から無作為抽出された匿名の人物のサンプルであった．2008年以降，個人のヒトゲノムがいくつか公表されている．DNA の 2 重らせん構造を発見したワトソン博士，ヒトゲノム解読で著名なベンター博士，匿名のアジア人および韓国人等のゲノムである. 分析の結果, 予想以上に細かい違いが多く，人種間の格差は大きい．一方，人種内の多様性は小さくなる傾向にあると考えられるので，日本人の標準的なゲノムの整備が進んでいる．また，米英中を中心とする研究機関が共同で「1000人ゲノムプロジェクト」を進めている．匿名性を保証しながら，1000人のゲノムを解析することで，塩基配列と個体の多様性の関係について，さまざまな知見が得られると期待されている.

医学以外の応用も広がっている．たとえば，経済動物（ニワトリ，ウシ，豚，等）や経済植物（イネ，小麦，キュウリ，キャベツ，イモ，ぶどう，等）のゲノム解析は盛んである．収穫の多い作物に特徴的なゲノム配列を探索することは，品種改良へとつながる．また，冷害や乾燥，害虫などに強い作物のゲノムを解読して，悪環境に強い個体

の塩基配列の特徴を検出し，品種改良を行うことで，寒冷地や乾燥地帯など，これまで作物が育ちにくかった地域を農地とすることができる可能性がある．

エネルギーや環境問題へのアプローチもある．トウモロコシやサトウキビからエタノールを作るバイオ燃料は，次世代エネルギーとして期待されており，実際ブラジルでは普及している．ただし，原料のトウモロコシの高騰を招くなどの問題を孕んでいる．そこで代替エネルギー源となる生物が探されている．たとえば植物から燃料を醸造するには，酵母や菌類を介在させるが，エネルギー変換を行う酵素遺伝子をさまざまな生物のゲノムの中に探すことで，廃材やサボテンなど穀物ではない植物から，効率よく燃料を醸造する技術が生まれる可能性がある．そのため米国エネルギー省は NIH(National Institutes of Health)と並んで，ゲノム解読に力を入れている．

2. 当拠点公募型共同研究として実施した意義

(1) 共同研究を実施した大学名　東京大学
(2) 共同研究分野　ゲノム科学　並列計算
(3) 当公募型共同研究ならではという事項など

ベンチマークではない現実のゲノムデータを限られた時間内に解析するために並列処理を研究開発している点．

3. 研究成果の詳細

HA8000 512 ノードを利用した生物配列アラインメントワークフローの性能分析
（実施者　鴨志田　良和　田浦健次朗　）

解析対象データとして，1000 人ゲノムプロジェクトより公開されたヒトゲノムデータ（公開に関する倫理申請が受理されたデータでサイズは 308GB）を入手した．HA8000 の 565 ノードを利用し，英国サンガーセンターが開発したパーソナルゲノム解析の典型的なワークフロー（図１）を実行した．

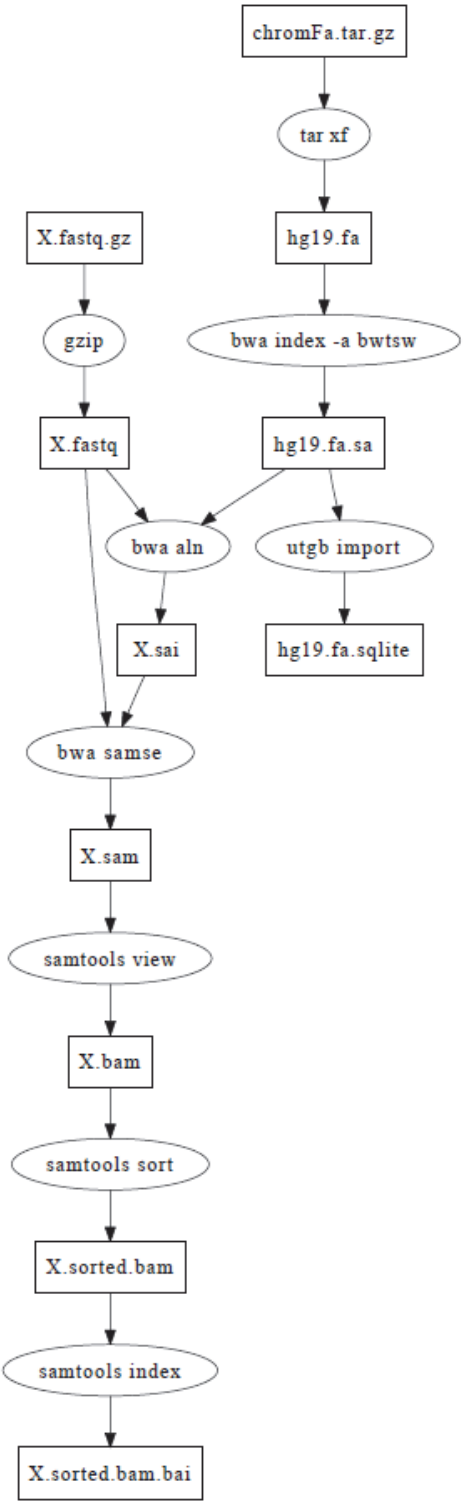

図１　1000 人ゲノム解析ワークフロー

一つの入力ファイルから 6 種類の中間/出力ファイルが生成され，そのファイル群のサイズ総合計は約 2.6TB である．またジョブは 565 ノードに割り振り実行した．

まず全体の実行時間であるが，図２に示すように最初の 1500 秒程度は 565 ノードがフルに並列稼

働していたものの，それ以降は 53 ノード程度の並列度であり同時に実行できるジョブの数は少なくなった．この結果から，十分に並列化されている部分もあるが，CPU コアを使いきれずワークフロー全体の終了時間を遅らせているジョブが混在している状況がうかがえる．ファイル I/O で必要な性能としては，2.6TB の中間ファイルを 1500 秒で書きだしたと考えると，平均 1.7GB/秒程度の I/O 性能があればよいことになる．

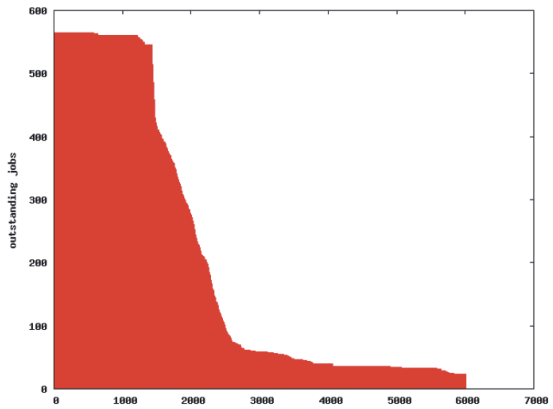

図２　横軸　実行時間　縦軸　並列稼働ノード数

図２に示す並列稼働ノード数を分析するために，ワークフロー中の各プロセスのノード内における CPU 稼働率を計測したのが図３～９である．図３に以下の６つのプロセスの CPU 稼働率を実行時間に沿って経時的に表示した．

- bwa aln
- gzip
- bwa samse
- samtools view
- samtools sort
- samtools index

bwa aln は接尾辞配列（以下慣習にならって suffix array）を利用して Burrows-Wheeler Transform を作成し，DNA 配列断片のゲノム上での位置を計算する（アラインメントする）ソフトウェア BWA を利用して計算を実行した結果である．DNA 断片が DNA 上に当たる位置は一般には複数あるため，複数の位置は suffix array 上の区間として表現され計算結果として返ってくる．bwa aln の CPU 稼働率を見ると（図４），良好に並列稼働していることがわかる．

つぎに巨大な配列データを圧縮した多数のファイル X.fastq.gz を解凍する gzip の CPU 稼働率は低い（図５）．これは複数のプロセスが同時にファイルシステムに負荷をかけるためにディスク I/O の遅延が生じてしまうことによる．今回の実験では，大量のアクセスに耐えるよう設計された並列ファイルシステム Lustre を使用している．しかし，このような遅延に対処するには，単純にジョブを並列に動かすだけでなく，各ジョブがファイルシステムにどのようなアクセスを行うかの情報を与え，ファイル I/O の性能限界を超えないようジョブスケジューリングの最適化を行う必要があることがわかった．

bwa aln が出力した suffix array 中のインデックスの区間は bwa samse により sam フォーマットと呼ばれる配列アラインメントを表現する標準フォーマットへと変換される．この CPU 稼働率も１未満であり（図６），並列化による性能向上の余地を残している．

bwa samse が出力するファイル X.sam は，samtools view によりさらにバイナリ形式である X.bam へと変換される．この手続きも並列化可能でありながら，実際のシステムではほとんど並列化されていないようである（図７）．

X.bam フォーマットではアラインメントが座標順に並んでいないため，座標に沿ったソーティングを行い整理する必要がある．これを実行するのが samtools sort である．ファイルが大きいため外部ソートを使った工夫が必要であるが，並列マージソートにより並列化が可能である．しかし実際には行われていないことは性能評価から明らかである（図８）．

最後に samtools sort が出力するファイル X.sorted.bam にインデックスを付与するのが samtools index である．CPU 稼働率からは並列化されていないようであるが（図９），実際，既にソートされたデータにインデックスを付ける処理なので，並列化の必要もなく短時間で実行が終了する．

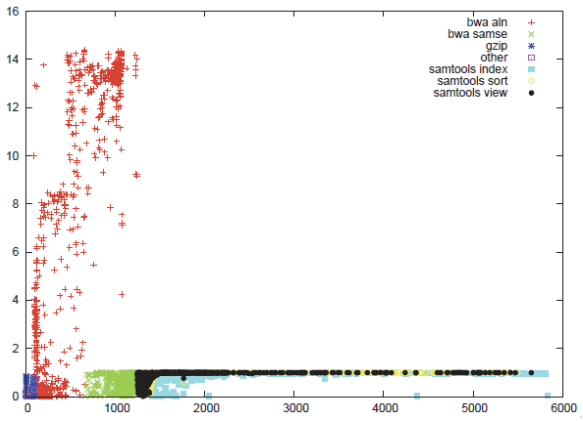

図 3　CPU 稼働率　(横軸：実行時間
縦軸：ノードあたり CPU 稼働率)

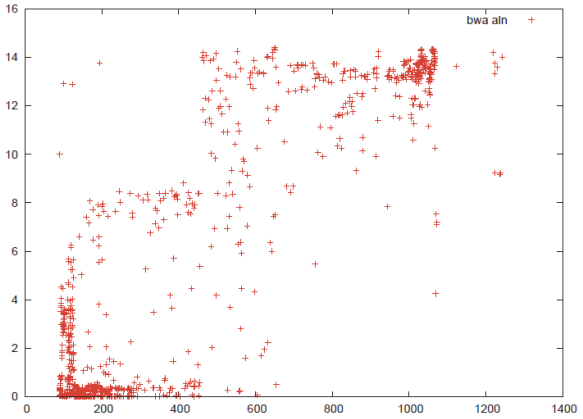

図 4　CPU 稼働率（bwa aln）
アラインメント

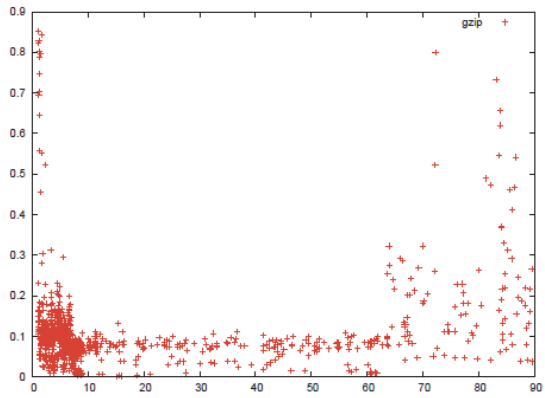

図 5　CPU 稼働率（gzip）

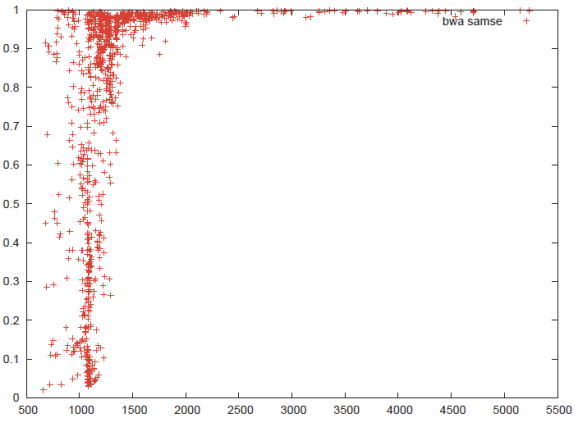

図 6　CPU 稼働率（bwa samse）

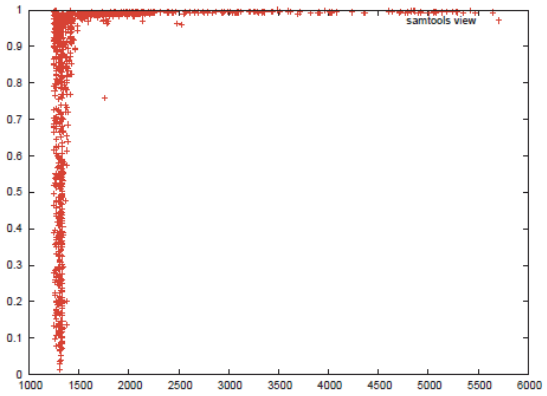

図 7　CPU 稼働率（samtools view）

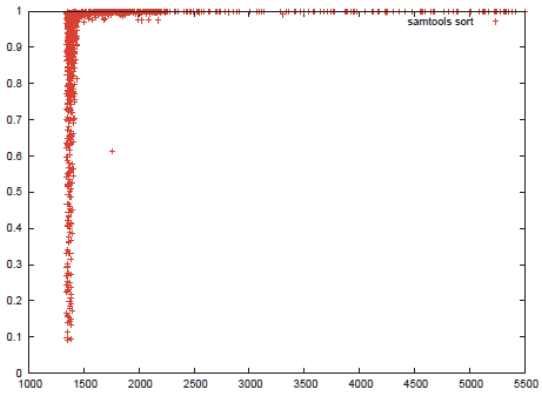

図 8　CPU 稼働率（samtools sort）

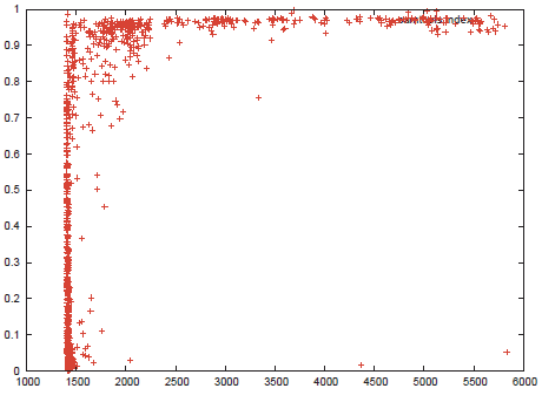

図 9　CPU 稼働率（samtools index）

以上のように，各プロセスにおける CPU 稼働率を吟味した結果，bwa samse，samtools view，samtools sort の３つのプロセスは効率化の障害となっており，しかもその遅延は並列化により解消できそうであることも理解できた．

大規模生物データ処理のための並列データベースシステム（実施者　斉藤太郎）

次世代シーケンサーの登場により，生物・医学の分野で扱うべき情報の量は飛躍的に増大したが，このような大規模データを扱うためのシステムはいまだ整っていない．例えば，ヒトゲノムの情報解析では，ヒトゲノムの標準参照配列(reference genome) と，シーケンサーで読まれた配列断片（read）を比較し，文字列比較によるアラインメントを実行するのが第一ステップとなる．その後，アラインメント後のデータを整理し，ゲノム配列中の遺伝子領域の情報，個人・集団特有の塩基の変異（SNP: Single Nucleotide Polymorphism），世界中の研究者が集めたアノテーション情報などを利用し，解明したい生命の謎，研究の目的に応じて，種々のデータを組み合わせた解析を行うことになる．

これらの膨大な情報処理を，ビジネス，ウェブ用途に進化してきた既存のデータベースシステム上で管理するのは非常に難しい．第一に，研究者間での情報共有を円滑に行うため，ゲノム関連の情報はすべてテキスト形式で管理されている．構造化したデータをもつテキストを，関係データベースシステム（RDBMS）のテーブルに変換する方法は自明ではなく，さらに，1TB に及ぶテキストデータを構文解析し，データベースのレコードに変換して保存(insert)するだけでも相当な時間を要する．本研究では，このようなテキスト処理，レコード変換のコストを減らすとともに，テキスト処理の並列化を行うための，生物データ処理に特化したデータベースシステムの設計を進めている．

また，種々の生物・医学データを組み合わせて計算するプログラムでは，アラインメントや統計解析など複雑な処理をともなうため，従来のデータベースのように SQL（関係データベース用の簡

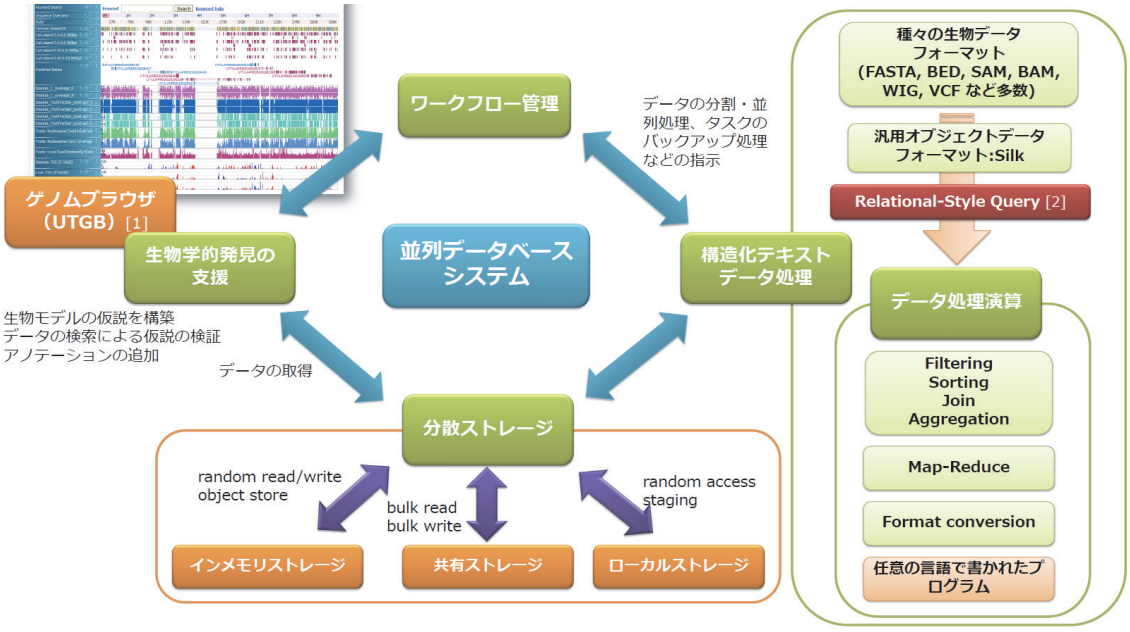

図１０　大規模生物データ処理のための並列データベースシステム

易な問い合わせ言語）で閉じた世界では，どうしても記述力が不足してしまう．そこで，任意のプログラミング言語からアクセスできるように，言語間の垣根を越えて利用できるデータベースシステムの設計を試みている．一方，先のワークフローの分析でも律速となっていた，データ処理（データのフィルタリング，ソーティング，テーブルの結合（join），MapReduce演算）など，あらゆるデータ操作の基本となるものは，並列・分散環境で高速に実行できるよう，各ノードが備えるCPUやコア数，メモリ量，共有・ローカルストレージの容量やLustre，HDD，SSDかなどの性能特性を配慮して並列化することが望ましい．近年のクラスタシステムでは，ノード当たりのメモリ量が数十GBを超えることも珍しくなくなってきたため，大容量メモリを活用したデータのソーティング，索引構築などの活用にも挑戦している．

さらに，入り組んだ生物データ処理のワークフローを確実に処理するため，ワークフローの並列実行，エラーのリカバリーや，テスト実行のための，ワークフローの部分実行(ワークフロークエリ）の開発にも取り組んでいる．

4. これまでの進捗状況と今後の展望

1000人ゲノムプロジェクトにおいて開発されたデータ解析ワークフローは実行効率の点で問題が多いことがわかった．われわれはボトルネックを洗い出し，並列化により解決できる見通しを得た．今後，我が国では，パーソナルゲノム解析が大幅に拡大し，医科学研究に波及してゆくことが考えられる．効率的なデータ解析を行い，疾患関連遺伝子を高速に感度よく描出してゆくため，図１０に示すプランを実現してゆくことに取り組む．

5. 研究成果リスト

(1) 学術論文（投稿中のものは「投稿中」と明記）

- Hongyan Wu, Taro L. Saito, and Shinichi Morishita. “Accelerating Path-free XML Queries in RDBMS.” IPSJ Transaction on Database. (in press)
- Taro L. Saito, Jun Yoshimura, Budrul Ahsan, Atsushi Sasaki, Reginaldo Kuroshu and Shinichi Morishita. Developing Personalized Genome Browsers with UTGB Toolkit. In Tag-based Approaches for Next generation Sequencing, Wiley-Blackwell-VCH (in press).

(2) 国際会議プロシーディングス

(3) 国際会議発表（招待講演）

- Shinichi Morishita. First International IEEE Conference on Computational Advances in Bio and medical Sciences (ICCABS 2011), Orlando, Feb 3-5 (2011)
- Shinichi Morishita. “Genetic Variation Associated with Nucleosome Structure and DNA Methylation.” Fish Genome Meeting 2011. Sanger Center, Cambridge. Mar. 11-12 (2011)
- Shinichi Morishita. “Genetic Variation Associated with Nucleosome Structure and DNA Methylation.” Biosoft 2011. Beijing. Mar 23-25 (2011)

(4) 国内会議発表

(5) その他（特許，プレス発表，著書等）

謝辞 本研究を進めるにあたり，データ分析を，田浦健次朗博士，鴨志田良和博士，斉藤太郎博士にお願いしました．ここに深謝します．またゲノムデータの収集については，文部科学省科学研究費新学術領域研究（研究領域提案型)『生命科学系３分野支援活動』「ゲノム科学の総合的推進に向けた大規模ゲノム情報生産・高度情報解析支援」，文部科学省科学研究費新学術領域研究（研究領域提案型)『パーソナルゲノム情報に基づく脳疾患メカニズムの解明』「脳疾患パーソナルゲノム多様性を分析する情報学の創成」の支援を受けています．またデータ解析については文部科学省グローバルＣＯＥプログラム「ゲノム情報ビッグバンから読み解く生命圏」の支援を受けています．