ゲノムワイド関連解析ソフトウェア PLINK 次期バージョン 1.90 における性能評価

金井 仁弘[1], 山根 健治[1,2], 樋口 千洋[1], 田中 敏博[1,3,4], 岡田 随象[1,5]

1. 東京医科歯科大学 大学院医歯学総合研究科 疾患多様性遺伝学分野　2. ソニー株式会社 メディカル事業ユニット 研究開発部門 LE 開発部　3. 東京医科歯科大学 疾患バイオリソースセンター
4. 理化学研究所 統合生命医科学研究センター 循環器疾患研究グループ　5. 理化学研究所 統合生命医科学研究センター 統計解析研究チーム

概要

米ハーバード大のPurcellらによって開発されたPLINKは、ゲノムワイド関連解析（GWAS）において広く利用されているソフトウェアである。我々は大幅な性能改善が謳われる次期バージョン1.90に対し、ベータ版を用いて具体的な処理性能を評価した。まずPLINKのサンプルデータ作成機能を用い、7段階のサンプル数（1,000～100,000）及びSNP数（10,000～1,000,000）からなる計49通りのジェノタイプデータを作成した。本データセットに対し異なるバージョンの PLINK（1.06, 1.07, 1.90b）を用いて、一般的なGWASデータに適応されるQC (quality control) 処理を実施した。その結果、1.90bは 1.06, 1.07 に比べて大幅な性能改善が認められ、大規模ジェノタイプデータに対する適合性が分かった。また、PLINK 1.90はソースコード共有ウェブサービスGitHub上でオープンソースとして公開されている。我々はこのGitHubを通じて、ベータ版に存在していたバグ除去に貢献したのであわせて報告する。

1. 背景

米ハーバード大のPurcellらによって開発されたゲノムワイド関連解析ソフトウェアPLINKの次期バージョン1.90 betaが登場した。

Version	Release Date	Developer
1.06	Apr. 24, 2009	S. Purcell
1.07	Oct. 10, 2010	S. Purcell
1.90b	**Jul. 1, 2014**	**S. Purcell, C. Chang**

計算機・次世代シーケンサーの性能が年々向上し、大規模データセットを扱う機会が増えた昨今、解析ソフトウェアの処理性能改善は研究時間の短縮に大きく寄与する。また適切な研究計画を練る上でも、各解析処理にどのくらいの時間を要するのか把握することが重要である。

2. 方法

一般的なGWASデータのQCプロセスで用いられる、以下6つのコマンドについて各バージョン (1.06, 1.07, 1.90b) での処理時間を測定した。同一データの下でのバージョン間の比較に加え、全測定データを用いてバージョン毎に処理時間とサンプル・SNP数の関係を計算した。またプロファイリングツールgprofを用いて、Genomeコマンドの処理を計測した。

1. コールレート `--missing`
2. ヘテロ接合度 `--het`
3. 連鎖不平衡 `--indep-pairwise`
4. IBS/IBD `--genome`
5. Hardy-Weinberg平衡 `--hardy`
6. アレル頻度 `--freq`

測定に用いたデータセットは1.90bのサンプルデータ作成機能を用いて生成されたサンプル数・SNP数、各7種類の計49通りのジェノタイプデータである。測定に用いたデータ・計算機の仕様は以下。

サンプル数	1,000　2,000　5,000　10,000　20,000　50,000　100,000
SNP数	10,000　20,000　50,000　100,000　200,000　500,000　1,000,000
CPU	Intel Xeon CPU E5-2450 v2 @ 2.50GHz × 16
Memory	96 GB

3. 結果　1.90bの処理時間は圧倒的に短縮されていた！

サンプル数 m：5,000・SNP数 n：100,000のデータセットにおける各処理とQCプロセス全体での所要時間をそれぞれ図1, 2に示す。1.90bは1.06, 1.07に比べて圧倒的に処理時間が短いことが分かる。QCプロセス全体（図2左）では、**1.90bは1.07に比べて約2,680倍速かった**（1.90b: 15秒、1.07: 11時間43分）。

特に時間を要するGenomeとIndepPairwiseを除いた比較(図2右)でも、1.90bは1.07に比べて約66倍速かった。また、全測定データを用いて処理時間 [s] とサンプル数 m・SNP数 n の関係を計算したところ図1上表を得た。この値にmnをかけると概ねの処理時間を得ることが出来る（Genomeのみm^2n）。

バージョン間での処理結果に丸め誤差以上の差異は認められなかった。

gprofを用いて、PLINKのGenomeコマンドの内部処理の様子を計測したところ図3を得た。PLINK 1.90bと1.07の間には抜本的な設計変更があるため単純な比較は出来ないが、やはりIBS/IBDの計算過程に大きな改善があることが分かる。また1.07はメモリ・文字列操作に関連する処理だけで1.90bの処理時間に匹敵する時間を要した。

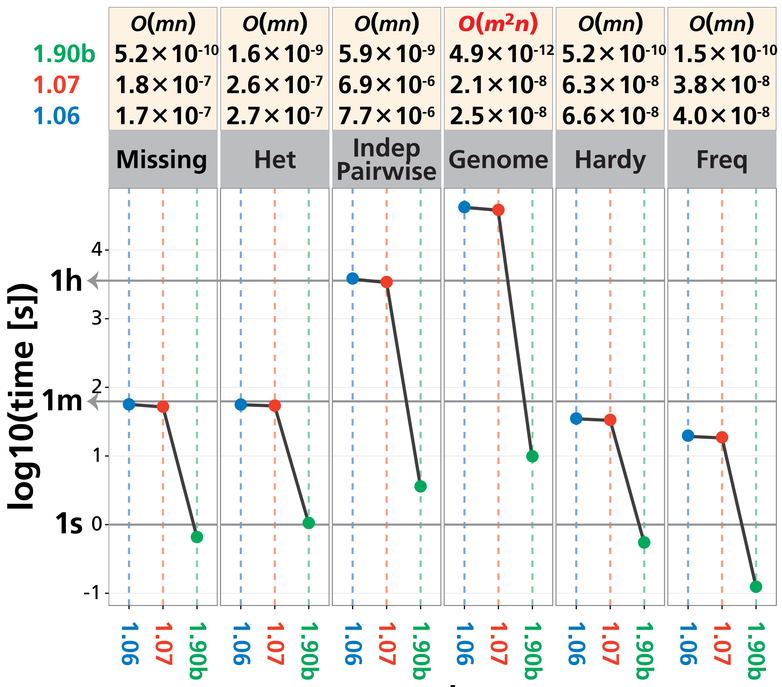

図1. 各コマンドの処理時間の比較
上：全測定データの基での比較　下：同一データの基での比較

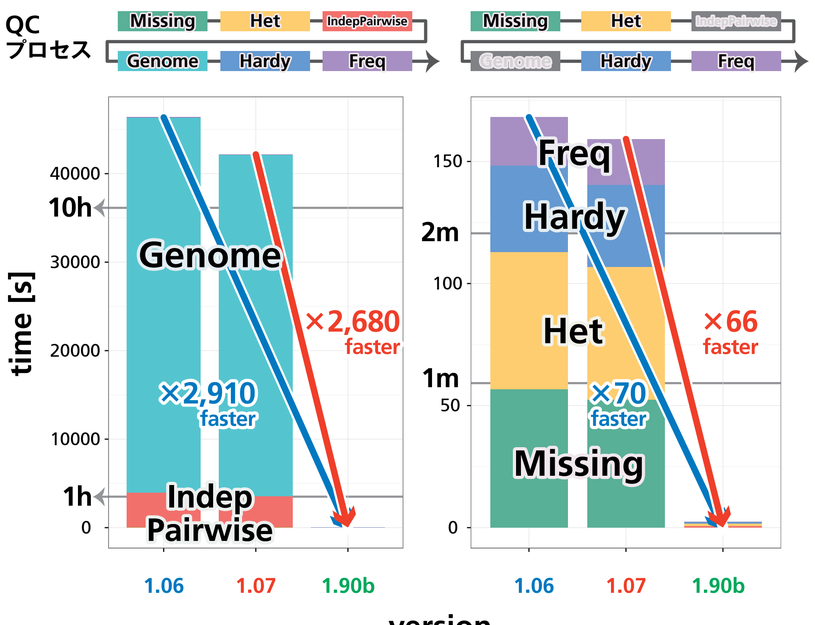

図2. QCプロセスの所要時間の比較
左：全コマンド　右：Genome・IndepPairwise以外

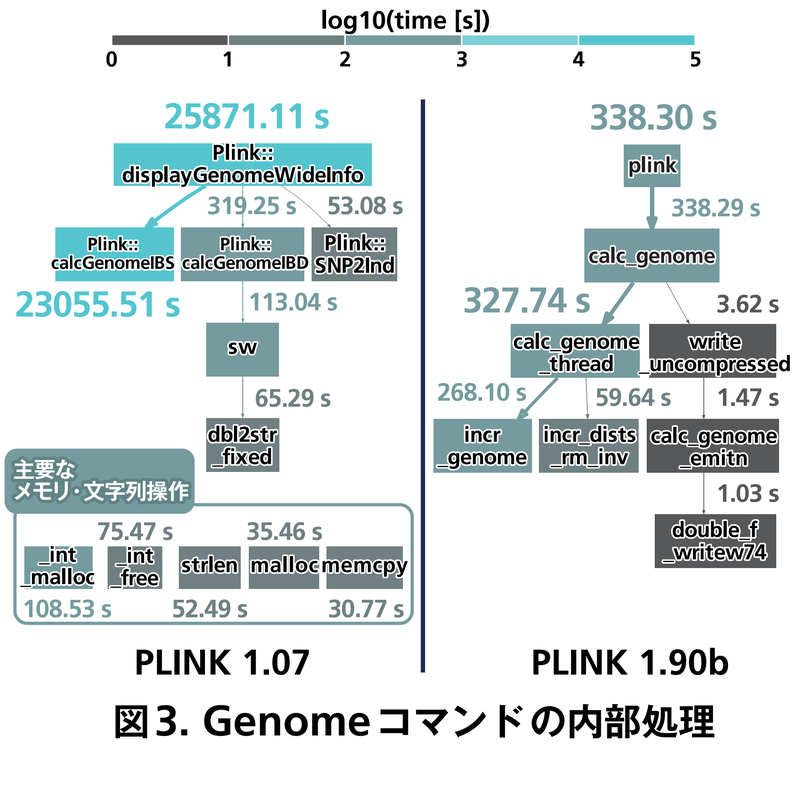

図3. Genomeコマンドの内部処理

4. ソースコードの解析

ソースコードレベルで確認すると、PLINK 1.90bで飛躍的に処理速度が向上した主な要因として、以下の3つが挙げられる。

1. bit演算や効率的なメモリアクセスといった抜本的な設計変更
2. アルゴリズムの改善・変更
3. 並列計算への対応（マルチスレッド、クラスター演算）

またgprofを用いた解析（図3）でも、これらの寄与が裏付けられた。

5. 結論

PLINKの次期バージョン1.90bの性能評価を行ったところ、QCプロセス所要時間の比較において**現行バージョン（1.06, 1.07）に比べ約2,500～3,000倍速い**ことが分かった。また今回測定した処理においては、1.90bと現行バージョンの出力の間に明らかな差異は認められなかった。

ソースコードの解析や実行時プロファイリングの結果から、1.90bには抜本的な設計変更やアルゴリズムの改善が施されており、これらが高速化に大きく寄与していることが確認された。

開発版への貢献：バグ修正

開発中のPLINK 1.9のソースコードが、GitHubを通じて公開されているため、開発者は自由にその設計を確認したり、機能の追加･修正を行ったりすることが出来る。

我々も開発版の以下の機能のバグについて修正パッチを作成し、本体に取り込まれた。

1. VCFパーサーのhalf-missing callの取り扱い
2. PEDパーサーの複数塩基・トリアレルの取り扱い
3. フェノタイプを含んだ共変量の出力

https://github.com/chrchang/plink-ng

連絡先：金井 仁弘　mkanai.brc@tmd.ac.jp