

統合データベース開発運用：多型知識表現技術開発  
 (我国における GWAS データの外部 QC とその結果の開示)  
 九州大学

1. 委託事業の9月末時点の判断基準になる目標
  - ・9月末には2個以上のGWASのQC結果をJAGQCで公開する。
2. 9月末時点の達成状況
  - ・4個のGWASのQC結果をJAGQCで公開した。
3. 上記達成状況を踏まえたプロジェクト終了までの目標
  - ・JAGQCパイプラインの拡充
  - ・未取得GWASデータの取得。

## 4. 成果の概要

データ請求件数：7 (うち3件はデータ未取得)

データ解析完了件数：4

解析項目：各取得データ当たり6項目(結果表示16細目)

サービス公開日：2010年9月30日(アクセスはデータ提供者に制限)

同10月6日～公開

アクセス件数：未集計(公開日数が短いため)

**JAGQC**  
Japanese Applied Genomics Quality Control

ホーム プロジェクト このサイトについて ENGLISH

**JAGQCとは**  
JAGQCは、我々の公的資金によって運行されているゲノムワイド関連解析研究(genome-wide association study: GWAS)で得られたデータの品質管理(quality check: QC)を第三者機関として一貫した統一的方法で行い、その結果をJAGQCデータベース(JAGQC DB)上に開示します。これによって公的資金によって行われているGWASの透明性を確保すると共に、GWASによって得られたデータを当初の関連解析以外の研究目的で再利用することを想定している外部研究者に、データの有用性に関する基礎情報を提供することを目的としています。

現在JAGQCで解析対象としているのはSNPを利用した大規模なケース・コントロールGWAS(ケース、コントロールそれぞれ約100人或以上、解析したSNP数300k或以上)です。対象とする各GWASにはJAGQCでのプロジェクト名が与えられます。

JAGQCで取得するのはSNPジェノタイプ(ディプロタイプ)データであり、各プロジェクトごとにJAGQCパイプライン(下記参照)によってそのQC解析を行います。取得したデータ及び個人情報を含む加工データはQC解析完了後JAGQCから抹消されます。従ってここで解析されたデータの利用を希望する研究者はそれらの情報元(各プロジェクトの履歴に記載)に直接連絡して下さい。

**JAGQC パイプラインの作業内容**  
取得したSNPジェノタイプデータの形式は必ずPLINK形式に変換されます。この際、含まれる試料IDはJAGQCでIIDに置換されます。また、取得したデータが2種類以上の解析プラットフォームから得られたものである場合は、各プラットフォームに共通したSNPのみについて解析を行います。QC解析は常染色体上のSNPについて、以下に述べる6個の解析ステップからなるJAGQCパイプラインで行います。ステップ1-ステップ5はPLINK<sup>1</sup>を、ステップ6はIGENSOFT/smartpca<sup>2</sup>を用いています。また、一部の統計解析、及び結果のグラフ表示には R を用いています(一部文庫3にあるスクリプトを利用)。解析結果はJAGQC DBにてログファイル、リスト(ダウンロード用圧縮ファイル)又はグラフとして開示されます。各SNPのD(rs#)及び染色体上の位置はデータ供給元からの情報に従っています。対応するdbSNP及びNCBI reference human genomeのバージョンは各プロジェクトの履歴に記載してあります(Figure 1)。

**Acquired data**  
Format/sample-ID conversion (shared SNP selection from all platform data)

**JAGQC pipeline**  
Step 1: Call rate qc of samples (sample call rate > 0.95)  
Step 2: SNP call rate qc of SNPs (SNP call rate > 0.95, maf > 0.02)  
Step 3: Missingness bias test (pair-wise r<sup>2</sup> > 0.20 (LD-pruning), H-W Equilibrium test)  
Step 4: Relatedness test by IBD score  
Step 5: Eigen decomposition  
Step 6: Stratification test by PCA

**JAGQC DB**  
-List of samples with missing rate  
-Graph of cumulative frequency  
-Logfile with # of samples at call rate < 0.95  
-List of SNPs with missing rate  
-Graph of cumulative frequency  
-Logfile with # of SNPs at call rate < 0.95  
-List of SNPs with no-call bias (p < 10<sup>-7</sup>)  
-Logfile  
-List of SNPs deviated from HWE (p < 10<sup>-7</sup>)  
-Logfile  
-Graph of p-hat distribution with # of sample pairs at p-hat = 0.1875  
-Logfile  
-Eigen vector values  
-Scatter plot of p1 vs. pc2  
-Logfile

このサイトについて  
Ver. 0.91  
Oct. 5, 2010 (0006)  
Minor editorial corrections and public release  
Ver. 0.9  
Sep. 30, 2010 (0530)  
Launch

ホーム プロジェクト このサイトについて

**プロジェクト一覧 (2010年10月1日現在)**

JAGQCでのプロジェクト	公開しているデータベース	JAGQCでの状況
Alzheimer's disease	GeMDBJ	解析完了
Pancreatic cancer	GeMDBJ	解析完了
Bronchial asthma	GeMDBJ	解析完了
Cerebral aneurysm	GWAS DB	解析完了
Hepatitis	GWAS DB	データ請求中
Narcolepsy	GWAS DB	データ請求中
Panic disorder	GWAS DB	データ請求中

各プロジェクトの対応するGWASに関する詳細についてはそれぞれを公開しているデータベースをご覧ください。

**Alzheimer's Disease (alz)**

履歴

公開しているDB (名称, [study id])	GeMDBJ, [Alzheimer's disease]
元となる研究組織 (名称, [代表研究])	ミレニアム・ゲノム・プロジェクト研究グループ (SGMGP) [不明]
ジェノタイプング実施施設	国立がん研究センター, Illumina San Diego
プラットフォーム	Illumina 550k/610k
公開しているDBへのデータ利用申請日	2010年6月22日
JAGQCでのデータ取得日	2010年8月17日
取得したファイル名	ALZ.CASE.txt, ALZ.CTRL.txt
試料数	763 cases, 1,422 controls
参照データベース (rs#, position)	dbSNP build 129, NCBI build 36

追加情報

SNP ポジション及びプラットフォーム (Illumina 550k [v.1, v.3]及びIllumina 610k) 間での共通SNPに関する情報はそれぞれ8月18日及び8月20日にGeMDBJから取得。共通するSNPに関するジェノタイプデータはJAGQCにて作成。JAGQCでは集計されたデータセット中のプラットフォーム間で共通するSNPについてqcを行う。プラットフォームごとにqcを行うことによりケース・コントロール間でコール失敗率に偏りのあるSNPは大幅に減少する可能性がある。JAGQCではこれを行っている。

解析結果

各試料のコール失敗率	alz_imiss.txt
試料コールレートの分布	alz_imiss.png

試料のコールレート