資料3－6

日本人ゲノム多型共有データベース

Shared database of **J**apanese **A**pplied **G**enomics

SJAG

(http://133.11.184.22/)

2007年1月8日

九州大学・生体防御医学研究所
林　健志

背景

超大量の個人ゲノムデータ生成とその管理法、解析法の目覚しい発展により、ゲノム配列情報を座標軸とした医科学が急速に進歩しつつある。

この発展にはデータの信頼度の明示が不可欠である。即ちデータ及び研究経過情報の迅速な共有と、これに続く多数の目によるquality controlとその結果のフィードバックが必須であり、これが基本的了解事項となっている。

< Lessons learned from HGP >

データ生産グループによる検証可能データ（生データ）の提供と、独立に存在するデータベースグループによる検証・公開が原則である。

解析材料提供によって情報学と生物学の接点を創成し、情報生物学者の質的・量的拡大を促進し、コミュニティのinformatics環境のレベルアップに貢献する。

SJAG

開設・運用：九大・林/日笠 (remote)

設置場所：ゲノム特定・基盤ゲノムタイピングセンター　(東大・医・人類遺伝、大量データ移行の利便性のため)

機械：Dell Linux server, terastation (my.sqlによるデータ管理、apache http serverでinternet接続)

現データ：上記センターで決定された「応用ゲノム・疾患関連解析」対照群試料のSNP頻度及びジェノタイプ情報。

[Affy 500K SNPs, 200人→1,000人]

SJAG に入ると

応用ゲノム共有データベース
(Shared database of Japanese Applied Genomics: SJAG)

version 0.3 (Jul, 2007)

- SJAGとは
- **SJAGの内容**
- File naming rule
- Directory構造
- ReadmeSJAG.pdf
- Level 1
- Level 2
- Level 3
- Scripts

このデータベースは利用者のinformatics 環境により、以下3種のレベルでの利用が可能です。

Level 1	**ウェブ・ブラウザーが使える**
	ウェブ・ページ "RSearcher"が起動します。これによってproject毎のアレル頻度、ジェノタイプ頻度、H-W平衡検定でのp値、各試料のジェノタイプ等をrs#又はSNP_A#によって検索(複数検索も可)することが可能です。各プロジェクトで決定された全試料での結果と、call rateによるfilteringを行った後の試料での結果を選択できます。H-W平衡検定は"SNP_HWE" (Wigginton JE et al., Am. J. Hum. Genet. 76: 887 - 893, 2005. [http://www.sph.umich.edu/csg/abecasis/Exact/])によるものです。更なるqcを必要とする可能性があるので目安と理解してご利用下さい。
Level 2	**解析プラットフォーム専用の解析ソフトウェアを利用可能**
	Affymetrix社500K のデータの場合はGCOS/GTYPEソフトウェア(英語版Windowsが必要)が必要です。各プロジェクトについて、ジェノタイピング(Nsp及び Styそれぞれのプロジェクト内全arrayデータを用いたBRLMM解析、confidence threshold = 0.50)を行った後、約32アレイごとに分割してexportした DTTファイル(アーカイブファイル、image fileを含まず)を、圧縮(.gz)して提供しています。
Level 3	**Perl scriptを用いて大量データの解析が可能(qc、CNV解析、新解析法のテスト等)**
	解析プラットフォーム専用の解析ソフトウェアの外で種々のデータ処理を行うことを想定しています。全てtext fileであり、必要に応じて圧縮(.gz または .zip)してあります。Affymetrix社500K のデータの場合は以下のファイルとなります。 □"Mapping250K_Nsp.na21.annot.csv" 及び "Mapping250K_Sty.na21.annot.csv"に各SNPのアノテーション情報(rs#、chromosome position、HapMapサンプルでのアレル頻度等)が記載されています。これらのファイルはAffymetrix社のサイトからの転載です。同サイトに最新版が出ている場合はそちらをお使い下さい。

Level 1 に入ると

RSearcher - Windows Internet Explorer

http://133.11.184.22/RSearcher.html

RSearcher (Genome Build 36, dbSNP Build 126)

Last Update 2007/06/13 by K.Higasa

Search by IDs

QC filter ⊙ ALL samples ○ Samples having call rate >= 0.95 only

Note: rs# and affy# must be prefixed with "rs" or "SNP_A-", respectively (i.e. rs233978, SNP_A-1780618)

Search Reset

Batch query for frequency

Upload the file 参照...

Format example

```
rs233978
rs9965312
rs10091369
rs2713901
rs17012816
rs5995963
```

QC filter ⊙ ALL samples ○ Samples having call rate >= 0.95 only

Search Reset

Hardy-Weinberg equilibrium test

Wigginton, J.E., Cutler, D.J., Abecasis, G.R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. American Journal of Human Genetics 76: 887 -- 893.

ページが表示されました インターネット 100%

複数検索の例　(call rate >= 0.95)

RSearch Results

Project ID	RS ID	Affy ID	Enzyme	Chr: position (strand)	Alleles	Genotype Count	Allele Frequency	Genotype Frequency	HWE (P-value)
301	rs233978	SNP_A-1780618	Nsp	4: 104894961 (+)	A/G	155	A : 0.303 G : 0.697	AA : 0.090 AG : 0.426 GG : 0.484	1.000e+00
301	rs9965312	SNP_A-1780617	Nsp	18: 24853920 (+)	C/T	155	C : 0.294 T : 0.706	CC : 0.097 CT : 0.394 TT : 0.510	5.612e-01
301	rs10091369	SNP_A-1780572	Nsp	8: 12915299 (-)	C/T	154	C : 0.185 T : 0.815	CC : 0.052 CT : 0.266 TT : 0.682	1.764e-01
301	rs2713901	SNP_A-1780357	Sty	15: 54398617 (-)	C/T	107	C : 0.079 T : 0.921	CC : 0.019 CT : 0.121 TT : 0.860	1.231e-01
301	rs17012816	SNP_A-1780352	Sty	4: 88667542 (+)	C/T	106	C : 0.330 T : 0.670	CC : 0.151 CT : 0.358 TT : 0.491	5.084e-02
301	rs5995963	SNP_A-1780351	Sty	22: 39632919 (-)	A/T	109	A : 0.927 T : 0.073	AA : 0.872 AT : 0.110 TT : 0.018	9.472e-02
301	rs5995963	SNP_A-1780351	Sty	22: 39632919 (-)	A/T	109	A : 0.927 T : 0.073	AA : 0.872 AT : 0.110 TT : 0.018	9.472e-02

直近の課題

0. 共有化の日程
1. 現タイピングセンターによる更なる対照群試料の解析
2. Affymetrix SNP Array 6.0 (含CNV)への対応
3. 他のプラットホーム（例：Illumina system）への対応
4. 「特定ゲノム」外からのデータ受け入れ
5. SJAG内でのdata qc、即ちcall rate, N/S concordance, H-W平衡, 層化（stratification)等を考慮
6. 疾患群データの取り扱い
7. 連結不可能匿名化試料での付随情報の取り込み
8. 外部大規模データ（海外：dbGap/FNIH/GAIN、Welcome Trust、国内：RIKEN/Biobank Japan）との協調？

長期展望

大規模ゲノム多様性データは拡大が必至。

10^6 SNPs x 10^2人 (現在)

→ 10^9 塩基 x 10^3人 (～3年後)

→ 10^9 塩基 x 10^8人 (10年後？)

これに対応する新しい管理・検証・解析・開示法の開発・運用には継続的な支援による情報技術の蓄積が不可欠である。