

「ライフサイエンス統合データベース開発運用」
(統合データベース開発:多型知識表現技術開発)

GWASデータ検証性促進のための QC Pipelineの作成とWEBページでの開示

九州大学生体防御医学研究所
林 健志

背景

疾患のゲノムワイド関連解析では、多数のcase(患者)群とcontrol(対照)群について $\sim 10^6$ 個のSNPに関するジェノタイプを決定し、両群間でアレル頻度に有意な(異常な)差があるSNPを検索し、その近傍に疾患要因となる遺伝的変異が存在すると結論する。

これが成立するのは

同質な集団からサンプリングされている？ 試料は互いに血縁関係がない？

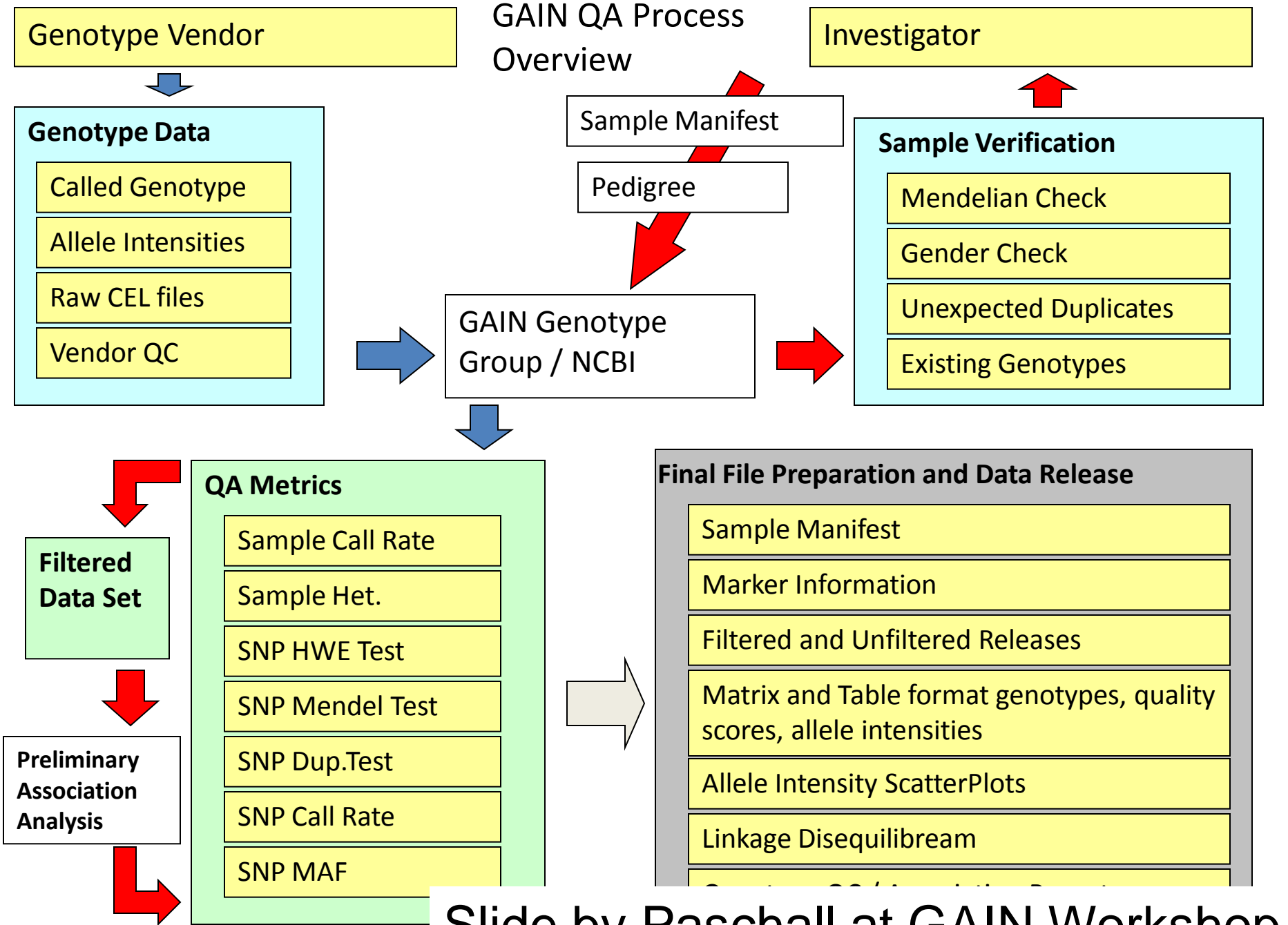
ジェノタイピングが正確？？？

等の条件をみたしている場合に限る。

現実のGWAS研究では、これらの前提は実現されていない。そこで得られたデータに対して種々のクオリティチェック(QC)を行い、前提からの逸脱の程度の多角的且つ定量的な検討とデータフィルタリングが必要。

GWAS Database はデータレポジトリとして多数の研究グループのGWASデータを扱う(データの質にばらつきがある)。データベースの質をどう維持するか？

先行研究(米国GAIN Project)ではどう対処しているか？



データQC機能を第三者に経由させる機構を統合データベースに組み込んでおくのが妥当。九州大学がこれを引き受ける(QCデータベースとして開示する)ことを提案している。

我国には統合データベース下のGWAS Databaseがあるが、このデータベースの運用機関と主たるデータ生産者が重複。

QCのためにはジェノタイプ生データ取得が必要。一方、第三者であるためには共同研究者であってはならない。

QCに作業を限定し、作業終了後は取得データを抹消することを前提。QC作業の内容を既定のQCパイプラインとする。全て事前に明記しておく。これによって倫理問題検討委員会からのデータ取得の承認が可能と考える。

結果をデータ提供元と協議。開示前に同意を得る。データを提供するか否かは各GWASプロジェクトのオプション。

JAGQCとは

本データベース (JAGQCdb) は、GWASにおける大規模ジェノタイプングデータの基本的なQCを一貫した方法 (JAGQC Pipeline) で行い¹⁾、そのプロセスと結果を加工せずに提示する。これによって疾患のゲノムワイド関連解析の検証性向上に役立てることを目的としている。なお、JAGQCで行われる作業は以下のQCに限定されており、受け入れたデータは本サイトから再配布されることはなく、また作業終了とともに当サイトから抹消される。

JAGQC Pipelineでは以下の2段階のQCを行い、それらの解析過程 (log file) 及び結果をダウンロード可能な形態で提示する。

一次QC: 受け入れたジェノタイプデータ²⁾に対してGAINQCプログラム³⁾を用いてsample QC、SNP QC、及びkinship QCを行う。またsample call rate、SNP call rateをもとにデータフィルタリングを行う。

二次QC: 一次QC/フィルタリングの後のデータセットに対してPLINKプログラム⁴⁾を用いて、ジェノタイプ失敗率の偏りの検定 (Test of missingness by case/control status)、集団の均一性の検討 (multidimensional scaling)、及び同プログラムのデフォルト設定での暫定的関連解析 (association) を行う。



¹⁾ 疾患のゲノムワイド関連解析では、多数 (数百人~数千人) のcase (患者) 群とcontrol (対照) 群について膨大な数の多型マーカー (~百万カ所のSNP) に関するジェノタイプを決定し、両群間でアレル頻度に有意な差 (基本的に帰無仮説に基づくカイ2乗検定で判定) があるSNPを検索する。そしてこのようなSNPの近傍に疾患要因となる遺伝的変異が存在すると結論される。この手法が単純に適用できるのは、両群が疾患要因以外は遺伝的に同質な集団からサンプリングされていること、試料は互いに血縁関係がないこと、ジェノタイプングが正確であり、且つタイプングの成功率 (sample call rate 及びSNP call rate) が十分高いこと、タイプングの失敗率 (missingness) に何らかのシステムティックな偏り (例えば両群間での) がないこと、等の条件をみたしている場合に限られる。しかしこれらの前提は実際のGWASでは必ずしも実現されていない。そこで現実のGWAS研究では、得られたデータに対して種々のクオリティチェック (QC) を行い、データセットの理想からの逸脱の程度を多角的且つ定量的に検討して、これに基づくデータフィルタリング等を経て、関連解析を行うことになる。また個々のGWAS研究は、広く受け入れられた手法によるデータQCの結果とその実行プロセスを明示して、関連解析の妥当性を主張する必要がある。

²⁾ Affymetrix社DNA Arrayによる解析に関しては出力生データ (.cel files) を入力データとし、本バイブライン内に組み込まれたAffymetrix Power Tool (APT)のBirdseed v2プログラムによってジェノタイプングが自動的に行われる。

³⁾ Abecasis et al.,
URL: <http://www.sph.umich.edu/csg/abecasis/GainQC/index.html>

⁴⁾ Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81, 559-575.
URL: <http://pngu.mgh.harvard.edu/~purcell/plink/>

Project List

ProjectID	Button
JNTM	 Go
JNTS	 Go

プロジェクトID: JNTM

データソースの課題名: 関連解析モデル1

研究代表:

プラットフォーム: Affymetrix SNP Array 6.0

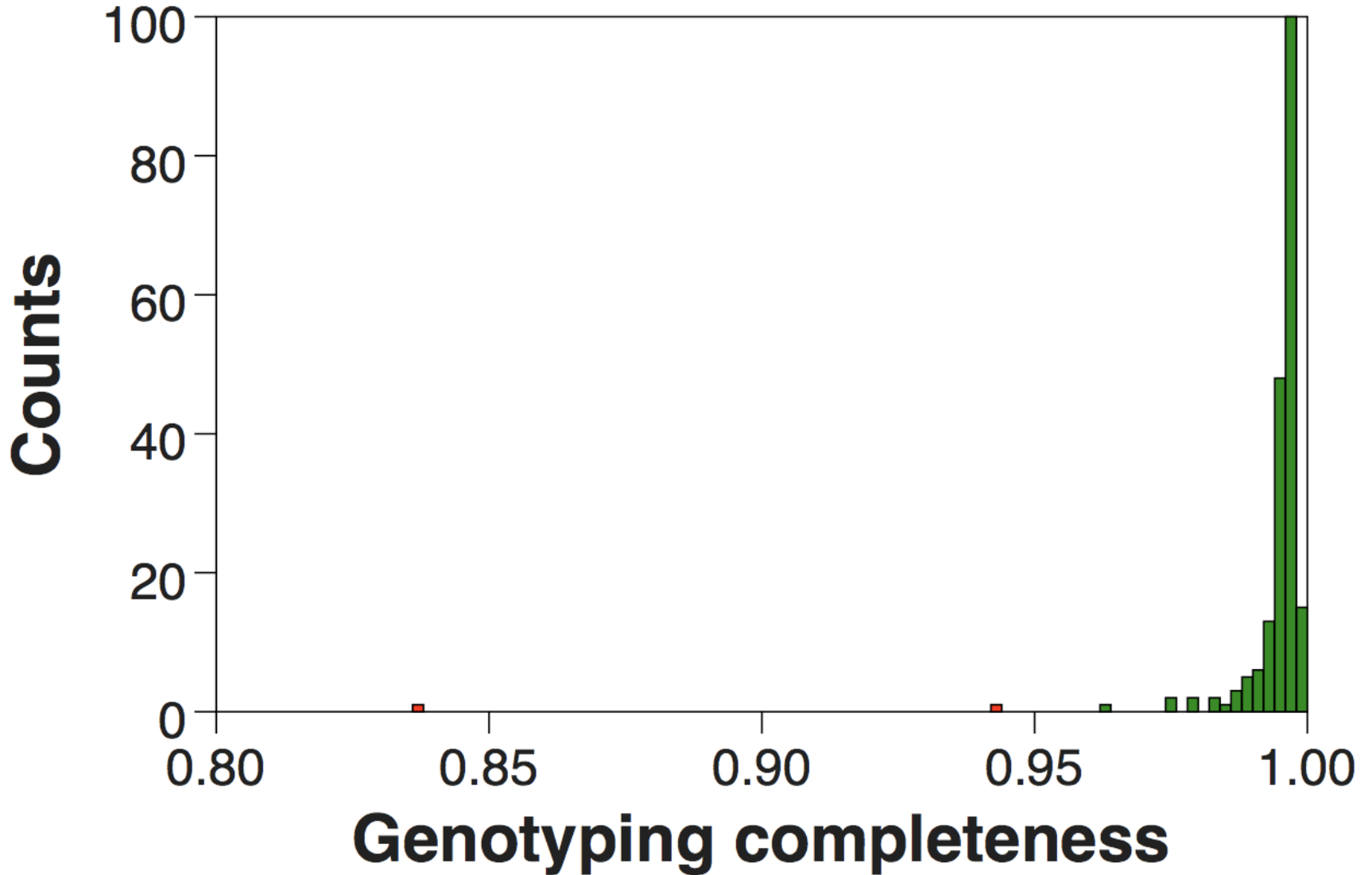
受け入れファイル: JNTM****.cel

URL: http://

GAINQC RESULTS

Description	File name	File size	Last Modified	Button
Summary	JNTM_GAINQC_SUMMARY.TXT	2kb	6-Jan-2009	↓ View
Sample Q.C	JNTM_GAINQC_SAMPLE.PDF	10kb	6-Jan-2009	↓ View
Snp Q.C.	JNTM_GAIN_QC_SNP.PDF	38kb	6-Jan-2009	↓ View
Relation Q.C.	JNTM_GAINQC_RELATION.PDF	3kb	6-Jan-2009	↓ View
Sample Table	---	---	20-Jan-2009	↓ View ↓ Download
Snp Table	---	---	20-Jan-2009	Chr: <input type="text"/> rs number: <input type="text"/> ↓ View ↓ Download
Relation Table	---	---	20-Jan-2009	↓ View ↓ Download

Sample genotyping completeness

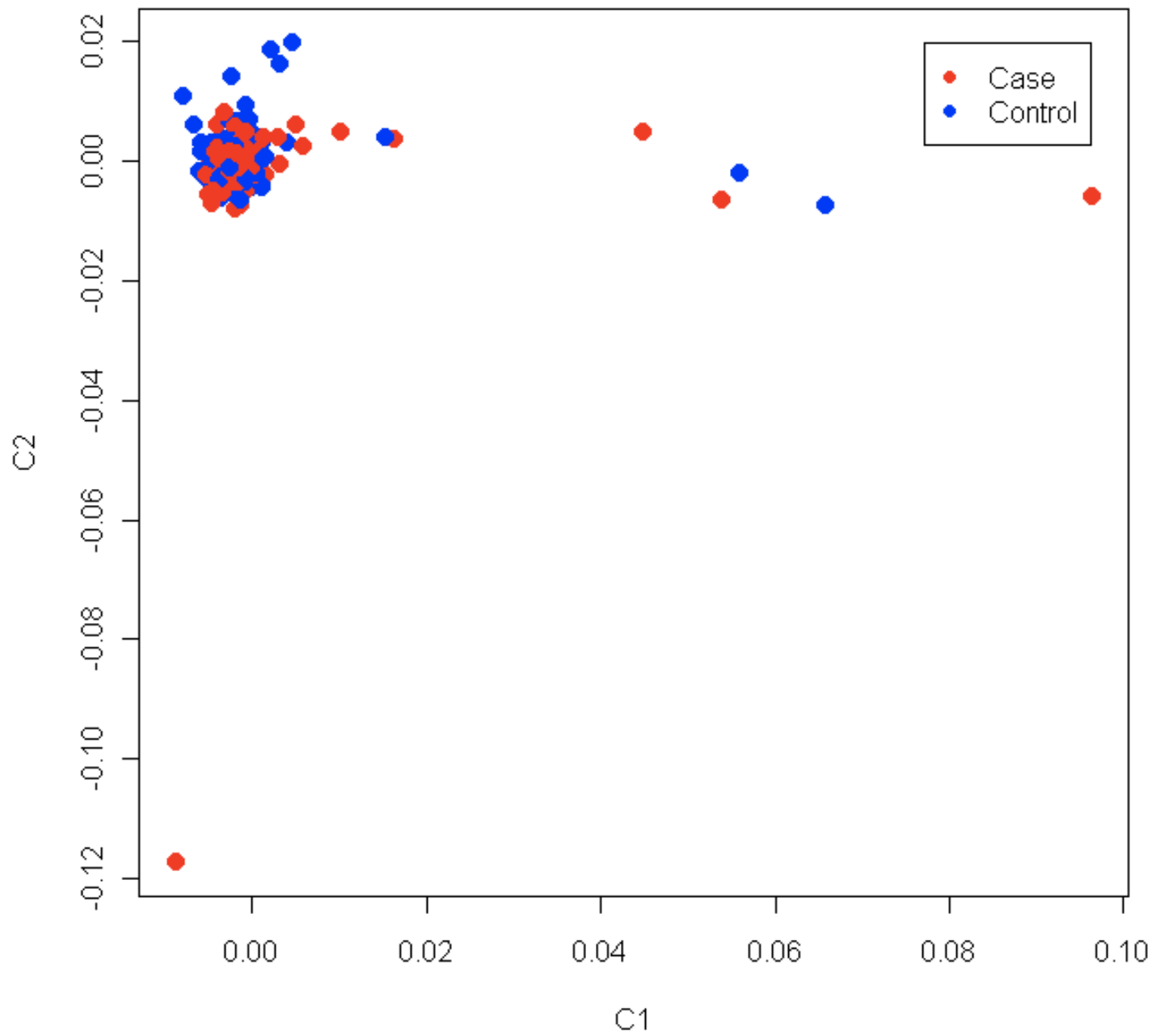


Sample information

id ▲	SampleId	Completeness	Heterozygosity	MendelErrors	SexOdds	LogL	AvgQualityScore	Flagged	Comments	FamilyId	PaternalId	MaternalId	Sex	Phenotype
1	JNT001	0.9962	0.2321	0	nan	nan	4819.4114	PASSED	-	JNT001	0	0	0	UNAFFECTED
2	JNT002	0.9964	0.237	0	nan	nan	4809.1957	PASSED	-	JNT002	0	0	0	UNAFFECTED
3	JNT003	0.9943	0.2384	0	nan	nan	4738.0436	PASSED	-	JNT003	0	0	0	AFFECTED
4	JNT004	0.9958	0.2379	0	nan	nan	4867.1951	PASSED	-	JNT004	0	0	0	UNAFFECTED
5	JNT005	0.9921	0.2398	0	nan	nan	4628.545	PASSED	-	JNT005	0	0	0	UNAFFECTED
6	JNT006	0.9942	0.2385	0	nan	nan	4790.2413	PASSED	-	JNT006	0	0	0	AFFECTED
7	JNT007	0.9956	0.2038	0	nan	nan	4866.8524	PASSED	-	JNT007	0	0	0	UNAFFECTED
8	JNT008	0.9946	0.2377	0	nan	nan	4813.0004	PASSED	-	JNT008	0	0	0	UNAFFECTED
9	JNT009	0.9932	0.2483	0	nan	nan	4698.0681	PASSED	-	JNT009	0	0	0	AFFECTED
10	JNT010	0.994	0.2355	0	nan	nan	4654.2846	PASSED	-	JNT010	0	0	0	UNAFFECTED
11	JNT011	0.9962	0.2364	0	nan	nan	4623.2137	PASSED	-	JNT011	0	0	0	UNAFFECTED
12	JNT012	0.9957	0.2371	0	nan	nan	4866.2162	PASSED	-	JNT012	0	0	0	AFFECTED
13	JNT013	0.9951	0.2381	0	nan	nan	4639.8328	PASSED	-	JNT013	0	0	0	UNAFFECTED
14	JNT014	0.9939	0.2358	0	nan	nan	4530.2665	PASSED	-	JNT014	0	0	0	UNAFFECTED
15	JNT015	0.9969	0.2355	0	nan	nan	4923.4594	PASSED	-	JNT015	0	0	0	AFFECTED
16	JNT016	0.9946	0.2385	0	nan	nan	4775.3082	PASSED	-	JNT016	0	0	0	UNAFFECTED
17	JNT017	0.9951	0.2444	0	nan	nan	4858.1078	PASSED	-	JNT017	0	0	0	UNAFFECTED
18	JNT018	0.9957	0.2456	0	nan	nan	4950.5229	PASSED	-	JNT018	0	0	0	AFFECTED
19	JNT019	0.9933	0.2381	0	nan	nan	4724.0062	PASSED	-	JNT019	0	0	0	UNAFFECTED
20	JNT020	0.9967	0.2356	0	nan	nan	5007.4931	PASSED	-	JNT020	0	0	0	UNAFFECTED
21	JNT021	0.9955	0.2345	0	nan	nan	4892.5782	PASSED	-	JNT021	0	0	0	AFFECTED
22	JNT022	0.9927	0.237	0	nan	nan	4687.1483	PASSED	-	JNT022	0	0	0	UNAFFECTED
23	JNT023	0.9928	0.2368	0	nan	nan	4747.2189	PASSED	-	JNT023	0	0	0	UNAFFECTED
24	JNT024	0.9917	0.2412	0	nan	nan	4653.8112	PASSED	-	JNT024	0	0	0	AFFECTED
25	JNT025	0.9941	0.2359	0	nan	nan	4703.4639	PASSED	-	JNT025	0	0	0	UNAFFECTED
26	JNT026	0.9936	0.2368	0	nan	nan	4669.8623	PASSED	-	JNT026	0	0	0	UNAFFECTED
27	JNT027	0.9936	0.2363	0	nan	nan	4560.0206	PASSED	-	JNT027	0	0	0	AFFECTED
28	JNT028	0.9725	0.2575	0	nan	nan	3954.7243	PASSED	-	JNT028	0	0	0	UNAFFECTED
29	JNT029	0.9928	0.2458	0	nan	nan	4815.2385	PASSED	-	JNT029	0	0	0	UNAFFECTED
30	JNT030	0.995	0.2353	0	nan	nan	4799.2766	PASSED	-	JNT030	0	0	0	AFFECTED
31	JNT031	0.9949	0.2479	0	nan	nan	4873.8789	PASSED	-	JNT031	0	0	0	UNAFFECTED
32	JNT032	0.9959	0.2442	0	nan	nan	4885.6328	PASSED	-	JNT032	0	0	0	UNAFFECTED
33	JNT033	0.9905	0.2452	0	nan	nan	4637.1697	PASSED	-	JNT033	0	0	0	AFFECTED
34	JNT034	0.9881	0.2455	0	nan	nan	4477.9599	PASSED	-	JNT034	0	0	0	UNAFFECTED

PLINKQC RESULTS

Description	Last Modified	Button
Sample(.imiss) table	10-Jan-2009	↓ View ↓ Download ↓ Log
Test of missingness by case/control status(.missing) table	10-Jan-2009	↓ View ↓ Download ↓ Log
Sample clustering	10-Jan-2009	↓ View ↓ Download ↓ Log
Assoc table	10-Jan-2009	↓ View .assoc ↓ Download .assoc.adjusted ↓ Download ↓ Log



Association test

id ▲	SNP	CHR	BP	A1	F_A	F_U	A2	CHISQ	P
1	SNP_A-8536815	4	35606525	G	0.08163	0.2222	A	15.08	0.000103
2	SNP_A-2057512	15	49203752	C	0.4495	0.2626	T	15.08	0.0001032
3	SNP_A-2236427	15	49157032	A	0.4545	0.2677	T	14.98	0.0001084
4	SNP_A-8375307	15	49166828	T	0.4545	0.2677	C	14.98	0.0001084
5	SNP_A-1791749	14	64131441	T	0.1263	0.2828	C	14.91	0.0001125
6	SNP_A-1803907	14	64153095	G	0.1263	0.2828	A	14.91	0.0001125
7	SNP_A-8461146	7	93388383	G	0.2828	0.1263	A	14.91	0.0001125
8	SNP_A-8342046	21	42705024	T	0.005051	0.08586	C	14.9	0.0001134
9	SNP_A-8342048	21	42716148	A	0.005051	0.08586	G	14.9	0.0001134
10	SNP_A-4194759	4	143999063	T	0.2194	0.08081	C	14.86	0.000116
11	SNP_A-8543335	4	35687788	T	0.1768	0.3469	G	14.77	0.0001213
12	SNP_A-8667660	7	13297268	C	0.101	0.2475	G	14.76	0.0001221
13	SNP_A-2237749	7	95242508	A	0.2908	0.1327	G	14.69	0.0001268
14	SNP_A-8420429	4	143985223	T	0.2677	0.1162	C	14.65	0.0001291
15	SNP_A-8593736	4	143995034	A	0.2677	0.1162	G	14.65	0.0001291
16	SNP_A-1845403	7	95242972	T	0.2879	0.1313	C	14.65	0.0001295
17	SNP_A-1856635	7	95242854	T	0.2879	0.1313	A	14.65	0.0001295
18	SNP_A-2215313	8	125285510	C	0.5758	0.3838	A	14.61	0.0001322
19	SNP_A-8333211	11	5031197	G	0.1443	0.3061	A	14.61	0.0001324
20	SNP_A-8670685	7	93383267	T	0.2857	0.1289	A	14.58	0.0001346
21	SNP_A-2035833	4	144118135	C	0.2316	0.08854	T	14.56	0.0001357
22	SNP_A-2158267	9	4084068	A	0.1061	0.2525	G	14.43	0.0001453
23	SNP_A-8569164	16	80070322	G	0.1364	0.2929	A	14.4	0.0001481
24	SNP_A-2158092	12	72069387	C	0.04545	0.1616	T	14.39	0.0001484
25	SNP_A-8479303	15	61006143	A	0.04545	0.1616	G	14.39	0.0001484
26	SNP_A-4207994	2	35713252	G	0.1061	0.01515	A	14.37	0.0001501
27	SNP_A-8717213	10	71087400	G	0.2755	0.4596	A	14.35	0.0001517
28	SNP_A-1845096	5	121013969	A	0.0202	0.1162	T	14.35	0.0001519
29	SNP_A-1904705	12	3439663	A	0.3182	0.1566	G	14.28	0.0001572
30	SNP_A-8698069	20	50670380	T	0.2857	0.1313	C	14.25	0.0001602
31	SNP_A-8578796	21	37061502	C	0.3737	0.202	A	14.24	0.0001609

	OR	L95	U95	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
	0.3111	0.1688	0.5733	0.0002042	1	1	1	1	0.9139	1
	2.293	1.502	3.498	0.0002045	1	1	1	1	0.9139	1
	2.28	1.497	3.473	0.0002141	1	1	1	1	0.9139	1
	2.28	1.497	3.473	0.0002141	1	1	1	1	0.9139	1
	0.3664	0.2176	0.617	0.0002214	1	1	1	1	0.9139	1
	0.3664	0.2176	0.617	0.0002214	1	1	1	1	0.9139	1
	2.729	1.621	4.595	0.0002214	1	1	1	1	0.9139	1
	0.05405	0.007121	0.4102	0.0002231	1	1	1	1	0.9139	1
	0.05405	0.007121	0.4102	0.0002231	1	1	1	1	0.9139	1
	3.197	1.732	5.901	0.0002278	1	1	1	1	0.9139	1
	0.4042	0.2529	0.646	0.0002373	1	1	1	1	0.9139	1
	0.3417	0.1944	0.6004	0.0002388	1	1	1	1	0.9139	1
	2.681	1.602	4.488	0.0002473	1	1	1	1	0.9139	1
	2.781	1.626	4.756	0.0002514	1	1	1	1	0.9139	1
	2.781	1.626	4.756	0.0002514	1	1	1	1	0.9139	1
	2.674	1.599	4.474	0.0002521	1	1	1	1	0.9139	1
	2.674	1.599	4.474	0.0002521	1	1	1	1	0.9139	1
	2.179	1.458	3.256	0.0002569	1	1	1	1	0.9139	1
	0.3823	0.2313	0.632	0.0002573	1	1	1	1	0.9139	1
	2.704	1.604	4.557	0.0002612	1	1	1	1	0.9139	1
	3.102	1.7	5.66	0.0002632	1	1	1	1	0.9139	1
	0.3512	0.2017	0.6114	0.0002801	1	1	1	1	0.9139	1
	0.3811	0.2292	0.6336	0.0002852	1	1	1	1	0.9139	1
	0.247	0.1146	0.5326	0.0002856	1	1	1	1	0.9139	1
	0.247	0.1146	0.5326	0.0002856	1	1	1	1	0.9139	1
	7.712	2.262	26.3	0.0002887	1	1	1	1	0.9139	1
	0.4471	0.2938	0.6805	0.0002915	1	1	1	1	0.9139	1

```

@-----@
          PLINK!          |          v1.05          |          11/Dec/2008
-----
(C) 2008 Shaun Purcell, GNU General Public License, v2
-----
For documentation, citation & bug-report instructions:
  http://pngu.mgh.harvard.edu/purcell/plink/
@-----@

```

```

Web-based version check ( --noweb to skip )
Recent cached web-check found... OK, v1.05 is current

```

```

Writing this text to log file [ ut_bin.log ]
Analysis started: Tue Jan 6 21:39:14 2009

```

```

Options in effect:

```

```

  --bfile ut_bin
  --allow-no-sex
  --assoc
  --adjust
  --ci 0.95
  --out ut_bin

```

```

Reading map (extended format) from [ ut_bin.bim ]
868162 markers to be included from [ ut_bin.bim ]
Reading pedigree information from [ ut_bin.fam ]
198 individuals read from [ ut_bin.fam ]
198 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
99 cases, 99 controls and 0 missing
0 males, 0 females, and 198 of unspecified sex
Warning, found 198 individuals with ambiguous sex codes
Writing list of these individuals to [ ut_bin.nosex ]
Reading genotype bitfile from [ ut_bin.bed ]
Detected that binary PED file is v1.00 SNP-major mode
Before frequency and genotyping pruning, there are 868162 SNPs
198 founders and 0 non-founders found
Total genotyping rate in remaining individuals is nan
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 868162 SNPs
After filtering, 99 cases, 99 controls and 0 missing
After filtering, 0 males, 0 females, and 198 of unspecified sex
Writing main association results to [ ut_bin.assoc ]
Computing corrected significance values (FDR, Sidak, etc)
Genomic inflation factor (based on median chi-squared) is 1.09348
Mean chi-squared statistic is 1.01327
Correcting for 699040 tests
Writing multiple-test corrected significance values to [ ut_bin.assoc.adjusted ]

```

```

Analysis finished: Tue Jan 6 21:40:24 2009

```


PLINKQC RESULTS

Description	Last Modified	Button
Sample(.imiss) table	10-Jan-2009	↓ View ↓ Download ↓ Log
Test of missingness by case/control status(.missing) table	10-Jan-2009	↓ View ↓ Download ↓ Log
Sample clustering	10-Jan-2009	↓ View ↓ Download ↓ Log
Assoc table	10-Jan-2009	↓ View .assoc ↓ Download .assoc.adjusted ↓ Download ↓ Log

今後の方針その1 : — QC database関連 —

1. GAINQC機能をPLINKQCへ移行する。
2. 更にQC機能を強化する(batch-qc, qq-plot他)。
3. 実際のGWASデータを受け入れて解析し、結果を開示する。

今後の方針その2：低頻度多型を含む日本人ハプロタイプデータベース構築作業の開始

<背景>

HapMapが提供しているアジア人ハプロタイプ情報はtrioを用いたEuropean、Africanと異なり、ディプロイドジェノタイプから推定されたものであり、数%の誤りがある。

我々は胞状奇胎(ヒトハプロイド細胞の集合体である)由来DNAを用いた日本人確定ハプロタイプマップを構築してきたが、約100検体からのデータであり、低頻度多型の情報が不十分である。

<そこで本研究では>

最近公開されつつある標準日本人数百人のディプロイドジェノタイプデータを取り入れて、低頻度多型情報までを含むハプロタイプマップを構築して、統合化DBの一部としての新たなデータベースとして公開するプロジェクトを開始する。

このハプロタイプマップは、現在最高水準にあるとされるHapMap3と比較しても2倍以上の試料数を用い、また胞状奇胎由来の誤りのないハプロタイプ情報をもとに決定されるので、アジア人ゲノムに関する極めて精度の高い情報を提供することになる。

このデータベースにはLD-bin情報、tagSNP情報、コピー数多型(CNV)情報を含める予定であり、我が国の疾患関連解析の精度を飛躍的に高める情報基盤となると想定している。