

「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データ検証性促進と GWAS情報基盤の強化

九州大学生体防御医学研究所
林 健志

九大 ①多型知識表現技術の開発

- 第三者によるQCが必要である。
- データベース統合化に必須な貢献がない。
- 多型データのQCは重要。引き継ぎは東大と一体で考えるべき。
- QCパイプライン等の開発は結構であるが、それがどの程度データ生産者に認知され、当初の目標である「九大を経由させるしくみ」が機能するかが判断のポイント。
- 研究課題として競争的資金でやるべき内容かもしれない。
- 各種GWASDBごとにQCが必要で信頼性を維持する必要がある。
- 九大を経由したからといって解析結果の透明性を確保したことにならない。
- 東大Gや他のSNPデータベース構築者との連携をうまく作れているのかがわからない。
- 東大Grの配布資料では、九大によるQCへの言及が無かったのは何故か？

背景

- 疾患のゲノムワイド関連解析 (GWAS) は、試料収集、ジェノタイピング操作、データ処理の全てのステップで**技術的にチャレンジング**な研究である。
- 従って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、多角的且つ定量的なデータの検討とデータフィルタリング等を行うことが必要。
- GWASのデータベースはデータレポジトリとして多数の研究グループのデータを扱うので、データベースの質をどう維持し、**信頼度を主張**するか？
- 先行研究 (米国 **GAIN Project**) では**データ生産者とデータベース管理者を分離**し、生データがデータ生産者からデータベースに提供される。
- データベース管理者はこれをもとに種々のデータQCを**公開された方法**で行うことで、GWAS研究の客観性、透明性を確保し、信頼性が維持されている。
- 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。

そこで

- 個人情報保護の立場を維持しつつ、データQC機能を第三者としての九州大学に経由させる機構を統合データベースに組み込み、QC作業過程及び結果を開示する。これによりデータ及び解析結果の透明性をはかる。
- 作業はQCに限定し、提供された元データの再配布は行わない。
- QC作業は明示された既定のQCパイプラインにより行う。
- 九大の確定ハプロタイプデータベース情報を統合DBに提供して正確なSNP/CNVハプロタイプ情報を含む日本人標準ゲノムデータベースの構築に協力する。

方策

1. QC機能を強化する(PLINKQC, PCA 解析他を行う)。
2. 実際のGWASデータを受け入れる。
3. D-HaploDBのデータ提供。



Purpose of JAGQC

本データベース（JAGQCdb）は、GWASにおける大規模ジェノタイピ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。

QC Pipelineの内容

glossary

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、ジェノタイピングが正確であり、且つタイピングの成功率（sample call rate及びSNP call rate）が十分高いこと、タイピングの失敗率（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ピングを行った結果を用いる。

³⁾ Illumina社BeadsArrayによる解析に関してはBeadStudioデータ解析ソフトウェアによるSNPジェノタイピングを行い、その結果のジェノタイプデータを用いる。

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.

⁵⁾ Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190. doi:10.1371/journal.pgen.0020190.

一次QC

受け入れたジェノタイプデータ^{2), 3)}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）を行った結果を提示する。

二次QC

受け入れたジェノタイプデータの中でcall rate 99%以上のSNPを用いて主成分分析（principal component analysis: PCA）⁵⁾を行い、各試料の第一主成分および第二主成分に関する2次元展開図およびこの解析時のlog fileを提示する。

Project ID	Button
JNTM	↓ Go
JNTS	↓ Go
Illumina Hapmap	↓ 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

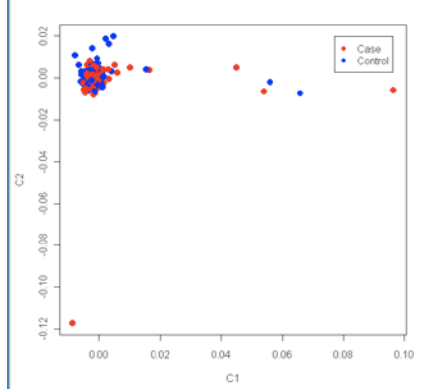
受け入れファイル: JNTM****.cel

URL: http://

PLINKQC RESULTS

Description	Last Modified	Button
Sample(.imiss) table	4-Jan-2009	↓ View
		↓ Download
		↓ Log
Test of missingness by case/control status(.missing) table	4-Jan-2009	↓ View
		↓ Download
		↓ Log
Sample clustering	4-Jan-2009	↓ View
		↓ Download
		↓ Log

FID	IID	MISS_PHENO	N_MISS	N_GENO	F_MISS
JNT001	JNT001	N	1751	868161	0.002017
JNT002	JNT002	N	1840	868161	0.002119
JNT003	JNT003	N	2760	868161	0.003179
JNT004	JNT004	N	1712	868161	0.001972
JNT005	JNT005	N	4025	868161	0.004636
JNT006	JNT006	N	2551	868161	0.002938
JNT007	JNT007	N	1929	868161	0.002222
JNT008	JNT008	N	2411	868161	0.002777
JNT009	JNT009	N	3113	868161	0.003586
JNT010	JNT010	N	2903	868161	0.003344
JNT011	JNT011	N	1767	868161	0.002035
JNT012	JNT012	N	1827	868161	0.002104
JNT013	JNT013	N	2456	868161	0.002829
JNT014	JNT014	N	3236	868161	0.003727
JNT015	JNT015	N	1177	868161	0.001356
JNT016	JNT016	N	2397	868161	0.002761
JNT017	JNT017	N	1928	868161	0.002221
JNT018	JNT018	N	1633	868161	0.001881
JNT019	JNT019	N	3271	868161	0.003768
JNT020	JNT020	N	1185	868161	0.001365
JNT021	JNT021	N	1766	868161	0.002034
JNT022	JNT022	N	3604	868161	0.004151
JNT023	JNT023	N	3626	868161	0.004177
JNT024	JNT024	N	4427	868161	0.005099
JNT025	JNT025	N	3012	868161	0.003469
JNT026	JNT026	N	3199	868161	0.003685
JNT027	JNT027	N	3389	868161	0.003904
JNT028	JNT028	N	18509	868161	0.02132
JNT029	JNT029	N	3177	868161	0.003659
JNT030	JNT030	N	2375	868161	0.002736
JNT031	JNT031	N	1993	868161	0.002296



```

@-----@
          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

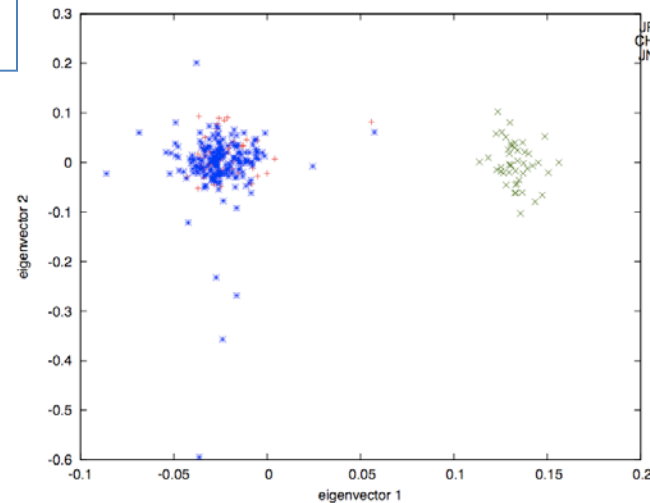
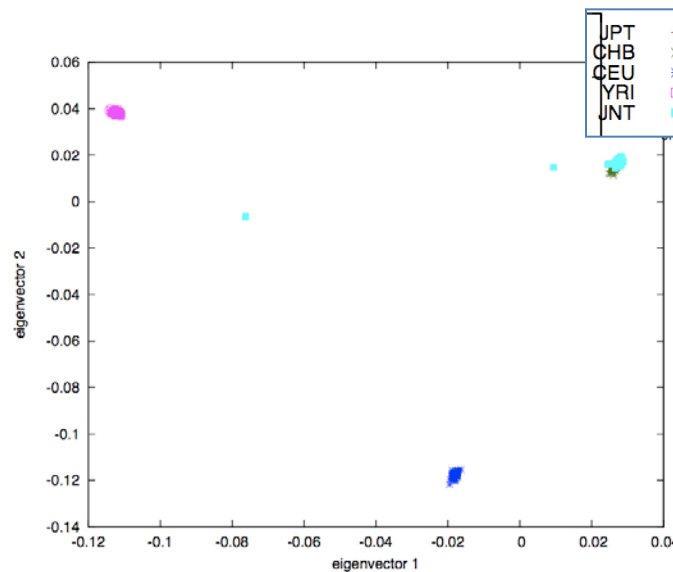
Reading map (extended format) from [ ut_bin.bim ]
868162 markers to be included from [ ut_bin.bim ]
Reading pedigree information from [ ut_bin.fam ]
198 individuals read from [ ut_bin.fam ]
198 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype 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

```

PCA RESULTS (before filtering)

Description	Last Modified	Button
CEU+YRI+CHB+JPT+JNT	8-Jan-2010	↓ View ↓ Log
CHB+JPT+JNT	8-Jan-2010	↓ View ↓ Log



```

parameter file: results/affy70c198/99w@haput2_OHG_JPT_NT198.ova.par
## IBC: NO MODE TESTS
PARAMETER NAME: VALUE
genotypefreq: total_99w@haput2_OHG_JPT_NT198.xmo
monoms: total_99w@haput2_OHG_JPT_NT198.xmo
indivname: total_99w@haput2_OHG_JPT_NT198.xmo
enccatnames: results/affy70c198/99w@haput2_OHG_JPT_NT198.ovc.evc
evcatname: results/affy70c198/99w@haput2_OHG_JPT_NT198.eval
allelestyle: NO
nummutvec: 6
nummut1refvec: 0
nummut1revvec: 3
outliers.genehash: 6.0
## amplicon version: 7530
norm used

genetic distance set from physical distance
number of samples used: 286; number of binsps used: 593369

## To get Tracy-Widom statistics: recompile matrix with TWSTAT correctly specified in README, or
just run twstats (see README file in POPGEN directory)

Kurtosis: ind
eigenvector 1 3.839 4.275
eigenvector 2 1.128 42.888
eigenvector 3 4.430 26.073
eigenvector 4 4.120 6.127
eigenvector 5 3.988 8.218
eigenvector 6 3.819 4.004
pooledancestrymap output
population: 0 JPT 45
population: 1 CHB 45
population: 2 UNT 198

## Average divergence between populations:
JPT CHB UNT p-value
CHB 0.007 0.007 0.0000
JNT 0.004 0.035 0.021
UNT 0.983 1.021 0.978 198

## Fat statistics between populations:
JPT CHB UNT fat std error
CHB 0.007 0.007 0.0000
JNT 0.007 0.007 0.0000

## Anova statistics for population differences along each eigenvector:
p-value
eigenvector_1_overall 0 +==
JPT minv -0.003 CHB maxv: 0.133
eigenvector_1_JPT_CHB +++
eigenvector_1_JPT_UNT 0.14676
eigenvector_1_CHB_UNT 0.0869621
eigenvector_2_overall 0.203920 JPT: 0.014
JPT minv -0.003
eigenvector_2_JPT_CHB 0.067131
eigenvector_2_JPT_UNT 0.0869621
eigenvector_2_CHB_UNT 0.0869621
eigenvector_3_overall 0.359216 CHB maxv: 0.006
JPT minv -0.003
eigenvector_3_JPT_CHB 0.017788
eigenvector_3_JPT_UNT 0.202694
eigenvector_3_CHB_UNT 0.026862
eigenvector_4_overall 0.782288 CHB maxv: 0.002
JPT minv -0.006
eigenvector_4_JPT_CHB 0.630002
eigenvector_4_JPT_UNT 0.630002
eigenvector_4_CHB_UNT 0.1987071
eigenvector_5_overall 0.614663 CHB maxv: 0.002
JPT minv -0.001
eigenvector_5_JPT_CHB 0.019594 JNT: 0.002
eigenvector_5_JPT_UNT 0.282542
eigenvector_5_CHB_UNT 0.05842
eigenvector_6_overall 0.781528 CHB maxv: 0.003
JPT minv -0.003
eigenvector_6_JPT_CHB 0.627891
eigenvector_6_JPT_UNT 0.627891
eigenvector_6_CHB_UNT 0.701745

## Statistical significance of differences between populations:
chi sq 5-value [post] [corr]
pop1 pop2 pop3
popdiff: JPT CHB JNT 232.978 6.945e-49 45 198
popdiff: JPT CHB JNT 8.862 0.715037 45 198
popdiff: JPT CHB JNT 77.234 2.463e-163 45 198

```


統合データベースプロジェクト ヒトゲノムバリエーションデータベース共有方針

現在、疾患関連遺伝子の探索においては非常な勢いでヒト遺伝子に関する情報が増加しており、これらの大量データの有効活用が急務である。文部科学省委託研究開発事業「統合データベースプロジェクト」はこのようなデータが広く共有されることを目的としており、様々なデータの共有に向けて施策を進めている。本方針は、「統合データベースプロジェクト・疾患解析DBグループ」（代表機関 国立大学法人 東京大学）（以下、「疾患解析DBグループ」）におけるゲノムワイドな多型情報および関連解析情報（GWAS）および疾患関連遺伝子のリシークエンシングデータについての共有方針について述べたものである。データの内容に応じて以下の6カテゴリーのデータを3つのレベルに分類して、それぞれのレベルに対する共有方針を定める。なお、データのデータベースへの提供とアクセスについては、「疾患解析DBグループ」内にデータ共有審査委員会（以下、「データ共有審査委員会」）を設置し審査を行う。委員会メンバーには「疾患解析DBグループ」の外部の委員を複数含めることとし、科学面・倫理面・その他の観点から検討を行う。

<本共有方針の概要>

データのカテゴリー／レベルとデータ使用時の手続き等について

レベル		1	2	3
データの 種類	GWASデータ	カテゴリーA（頻度データ、統計解析結果）	カテゴリーB（個体レベルでのCNV情報）	カテゴリーC（GWAS遺伝子型データ） カテゴリーD（GWAS生データ）
	リシークエンシングデータ	カテゴリーE（公知のリシークエンシングおよび変異データ）	カテゴリーF（リシークエンシングおよび変異データ）	（該当なし）
データ使用時の手続き等		ウェブサイトにおいて、閲覧可能。但しカテゴリーAのデータを大量取得する場合はレベル2と同様の申請を要する。	氏名、職名、連絡先、使用目的、e-mailアドレス（原則、所属機関から発行されたアドレス）を申請する。	データアクセス申請書を提出し、許可を受ける。また、使用期間に応じて、データ使用報告書の提出を要する。

データの安全な取り扱いについて

レベル3のデータに関しては、データアクセス同意書(http://gwas.lifesciencedb.jp/gwasdb/db_form/db_duc.pdf)を提出し、セキュリティガイドライン(http://gwas.lifesciencedb.jp/gwasdb/gwas_sec.html)に従って、アクセスの許可されたデータを安全に取り扱うものとする。

大量のレベル1データおよびレベル2データに関しては、ウェブサイト上に記載されたデータアクセス条件(http://gwas.lifesciencedb.jp/gwasdb/db_term.html)に同意するものとする。

日本人確定ハプロタイプ情報の統合DBでの導入

<背景>

関連解析での imputation、今後のarray設計はハプロタイプ情報に依存している。

HapMapのハプロタイプ情報: European、Africanは高精度、アジア人は低精度 (数%のphaseの誤り)。

D-Haploのハプロタイプ情報: 日本人胞状奇胎(ハプロイド細胞)由来DNA を用いており、高精度(日本人確定ハプロタイプ情報)。CNV情報がより確実。

Resequencingによって得られるハプロタイプ情報: 急速に拡大が期待されるが、低精度(数%のphaseの誤り)。D-Haploの情報によって精度向上が可能。

<そこで本研究では>

D-Haploの情報、標準日本人のディプロイドジェノタイプ情報等を統合して、低頻度多型情報までを含む高精度ハプロタイプマップを構築し、新たなデータベースとして公開するための方法を検討する。これはアジア人の疾患関連解析の精度を高める情報基盤となる。

Showing 200 kbp from Chr21, positions 14,580,000 to 14,779,999

Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

[\[Hide banner\]](#) [\[Bookmark this\]](#) [\[Link to Image\]](#) [\[High-res Image\]](#) [\[Help\]](#) [\[Reset\]](#)

Landmark or Region:

Search

D-Haplo Phase I and II, on NCBI Build35 Assembly

Reports & Analysis:

Annotate LD

Configure...

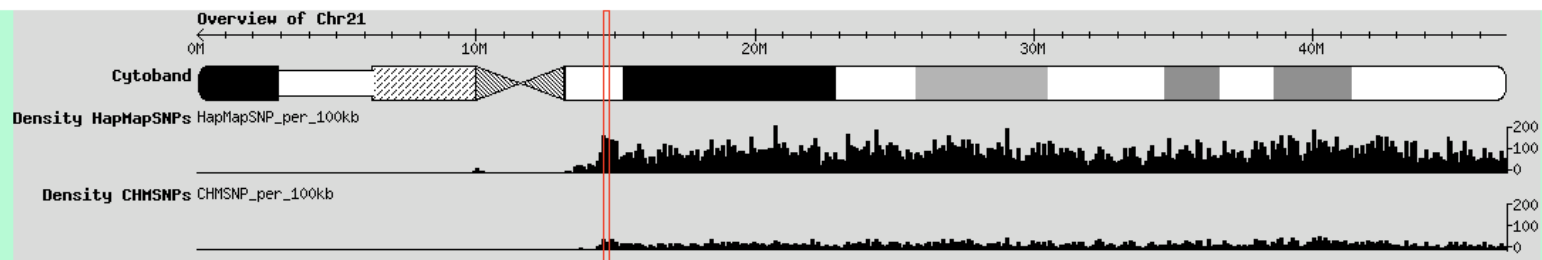
Go

Scroll/Zoom:

Show 200 kbp

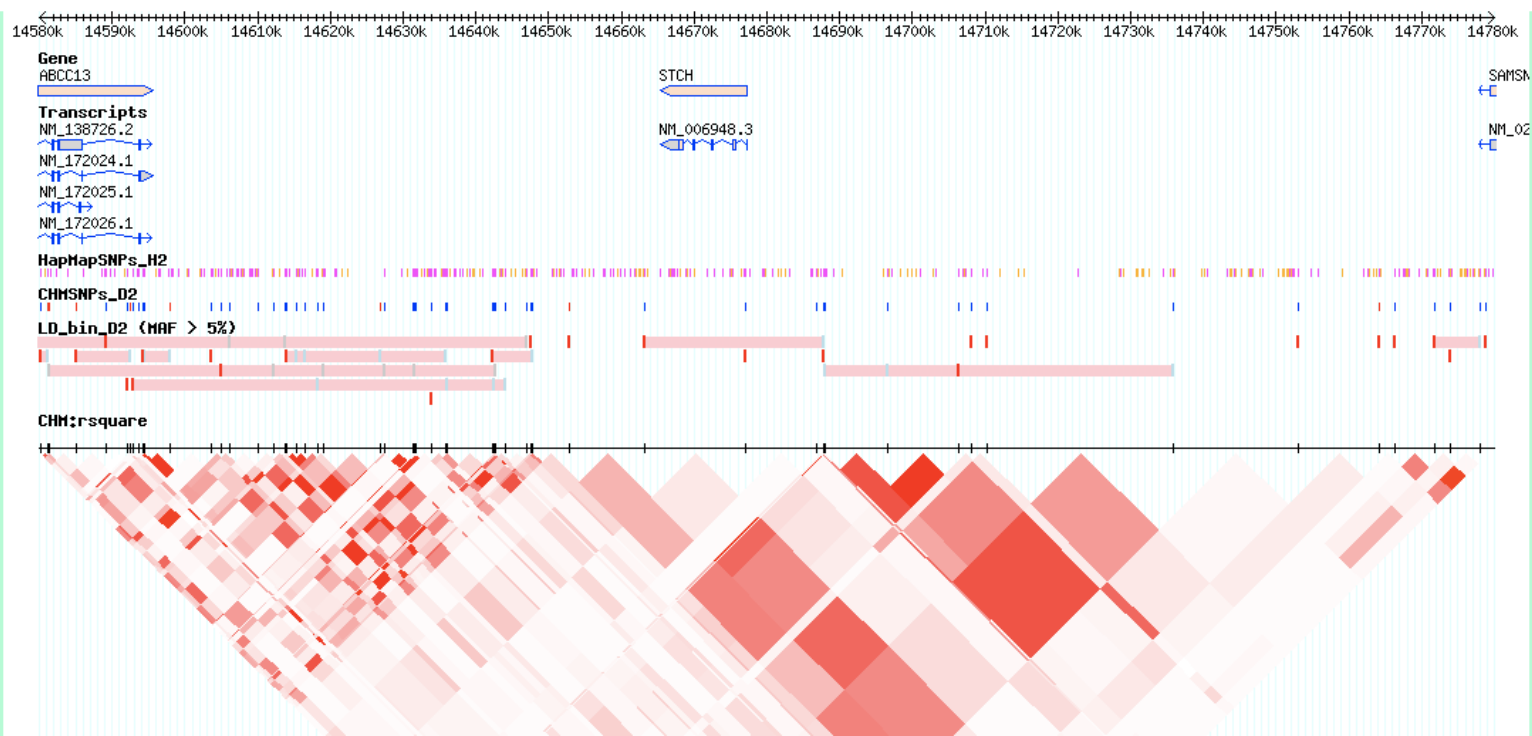
☐ Flip

[-] Overview



⊕ Region

Details



Mapping complex disease traits with global gene expression

By Cookson, Liang, Abecasis, Moffatt and Lathrop

FURTHER INFORMATION

Liming Liang's homepage:

<http://www.sph.umich.edu/csg/liang>

Abecasis laboratory homepage (contains programs for genome-scale data analysis):

<http://www.sph.umich.edu/csg/abecasis>

Catalog of Published Genome-Wide Association Studies:

<http://www.genome.gov/gwastudies>

Database of Genomic Variants:

<http://projects.tcag.ca/variation>

dbSNP: <http://www.ncbi.nlm.nih.gov/projects/SNP>

➔ **D-HaploDB:** <http://orca.gen.kyushu-u.ac.jp>

Genotype-Tissue Expression (GTEx):

<http://nihroadmap.nih.gov/GTEx>

H-InvDB: <http://www.h-InvDB.jp>

mRNA by SNP Browser v 1.0.1:

<http://www.sph.umich.edu/csg/liang/asthma>

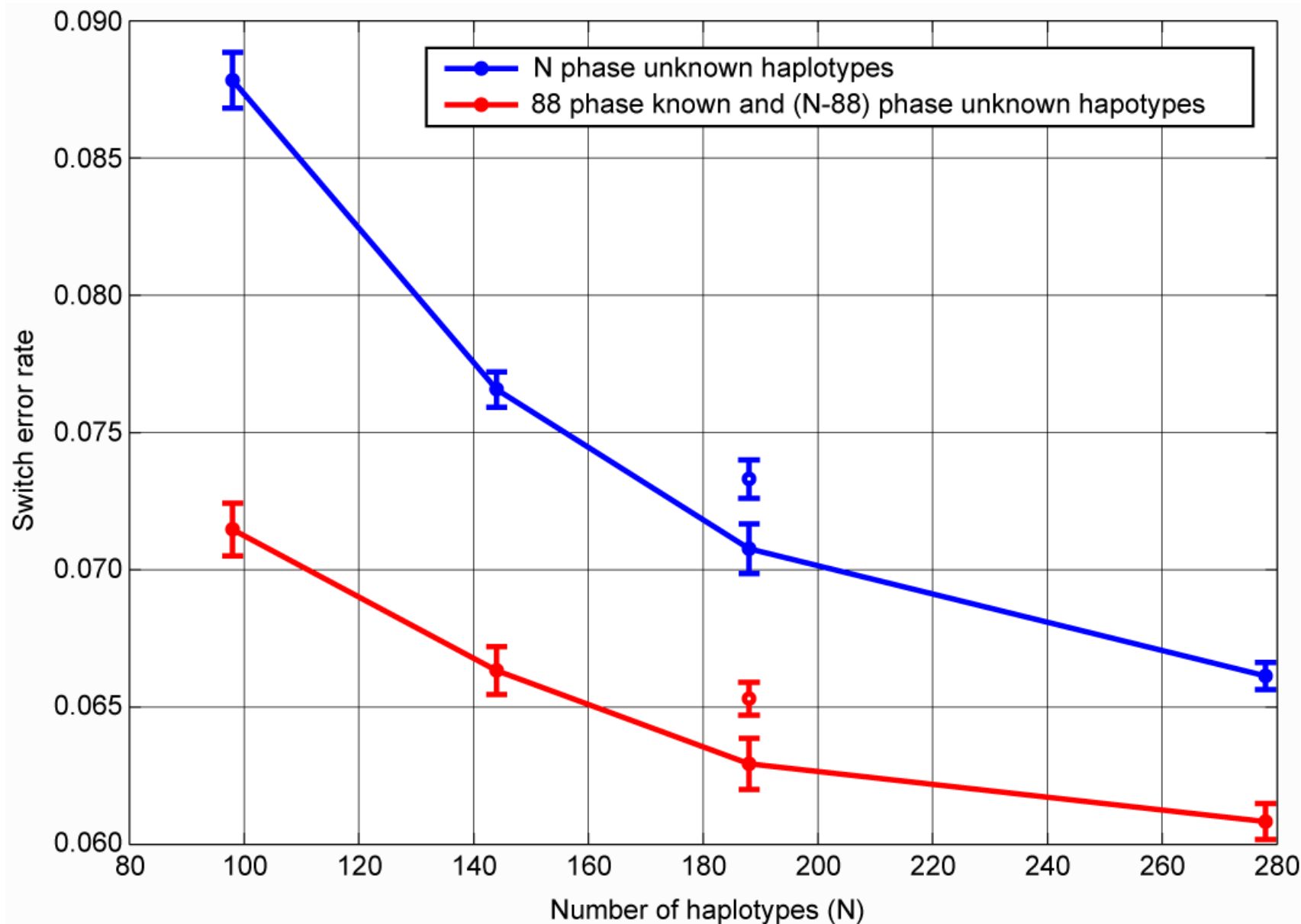
UniGene: <http://www.ncbi.nlm.nih.gov/unigene>

VarySysDB:

<http://www.h-InvDB.jp/varygene/home.htm>

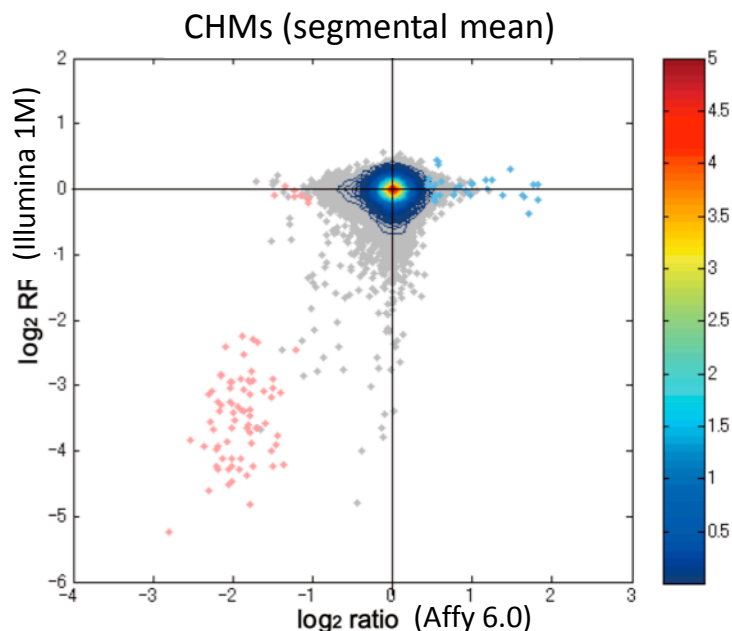
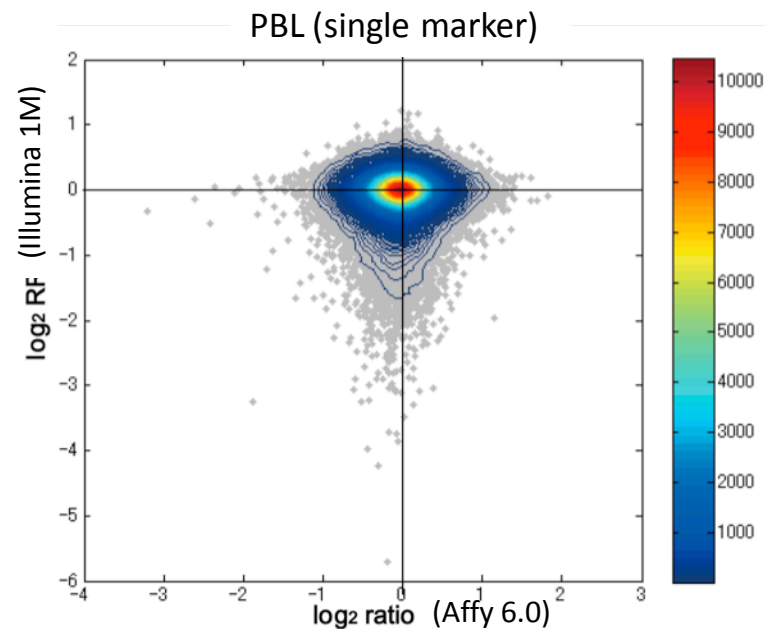
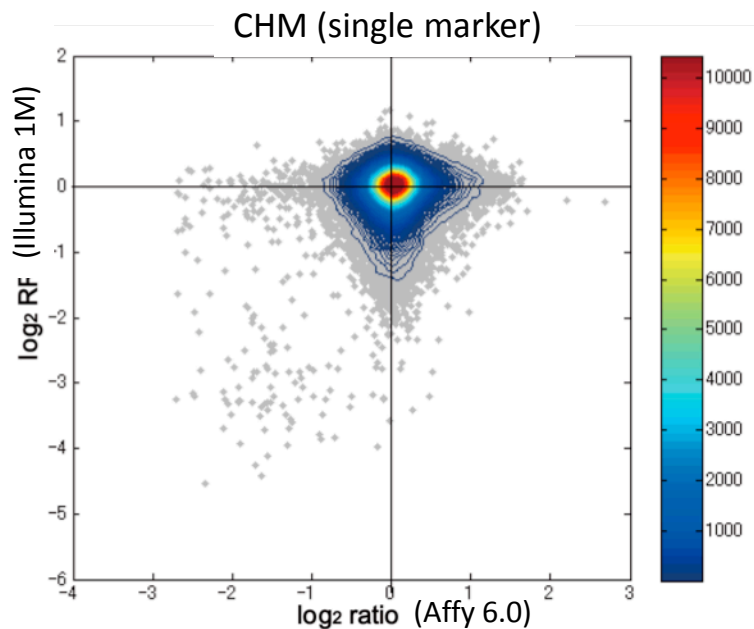
WaferGen: <http://www.wafergen.com>

D-Haplo情報によるディプロイドハプロタイプ推定精度の向上



(Higasa et. al., PlosGenet. 2009)

Haploid vs. Diploid in CNV Detection

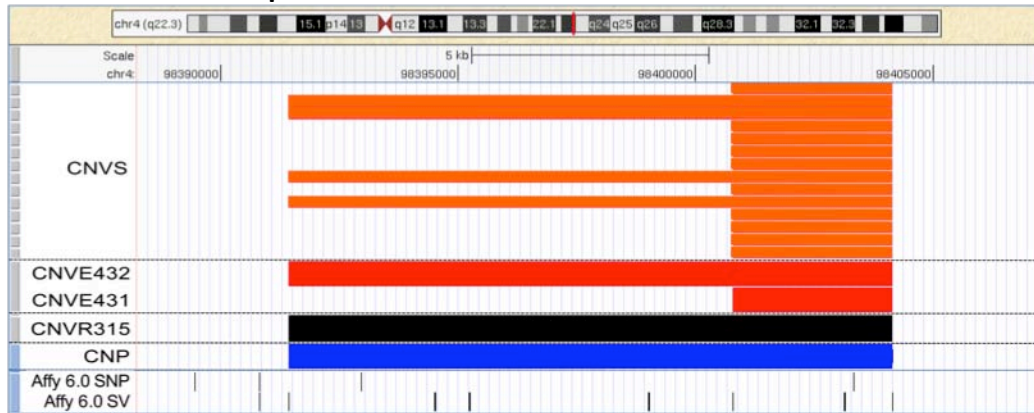


- In CHM (haploid sample), most CNV (deletion) markers were detectable as a cluster in the third quadrant, whereas in PBL (diploid sample), they were located at the tail portion of the distribution of normal copy markers.
- After segmental evaluation of CNVs, both the clusters of normal copy and deletion segments were detectable as distinct and compact clusters.

*Kukita et al., A.J.H.G. (2010),
doi:10.1016/j.ajhg.2010.05.003*

Haplotype Preference of Copy Number Variation Events

Deletion map of CNVR315 (chr4:98,191,000-98,605,000)



The American Journal of Human Genetics (2010),
doi:10.1016/j.ajhg.2010.05.003

“Definitive haplotype map determined by genotyping duplicated haploid genomes and finding of predominant haplotype preference of copy number variation events”

Yoji Kukita, Koji Yahara, Tomoko Tahira, Koichiro Higasa, Miki Sonoda, Ken Yamamoto, Kiyoko Kato, Norio Wake, Kenshi Hayashi

Haplotype patterns of 200 kb regions adjacent to CNVR315

