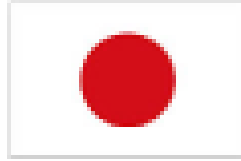
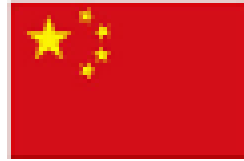


**The 9<sup>th</sup> Korea-Japan-China Bioinformatics Training Course**  
**Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences**  
2010年4月19日 ~ 2010年4月23日



# Genome-wide Association Study (GWAS)

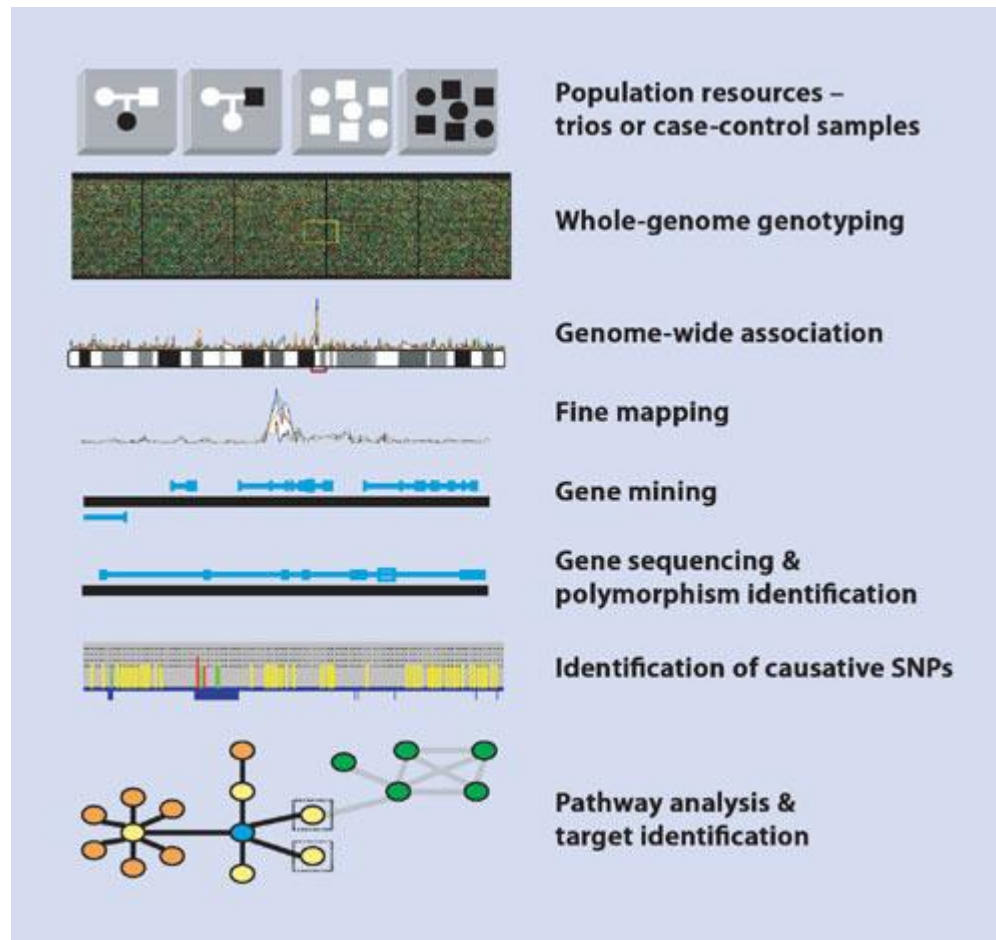
Sangsoo Kim

School of Systems Biomedical Sciences  
Soongsil University, Seoul, Korea

# What is GWAS?

- An examination of genetic variation across a given genome
- Designed to identify genetic associations with observable traits
  - Such as blood pressure or weight,
  - or why some people get a disease or condition
- Hypothesis-free approach
  - Candidate gene approach

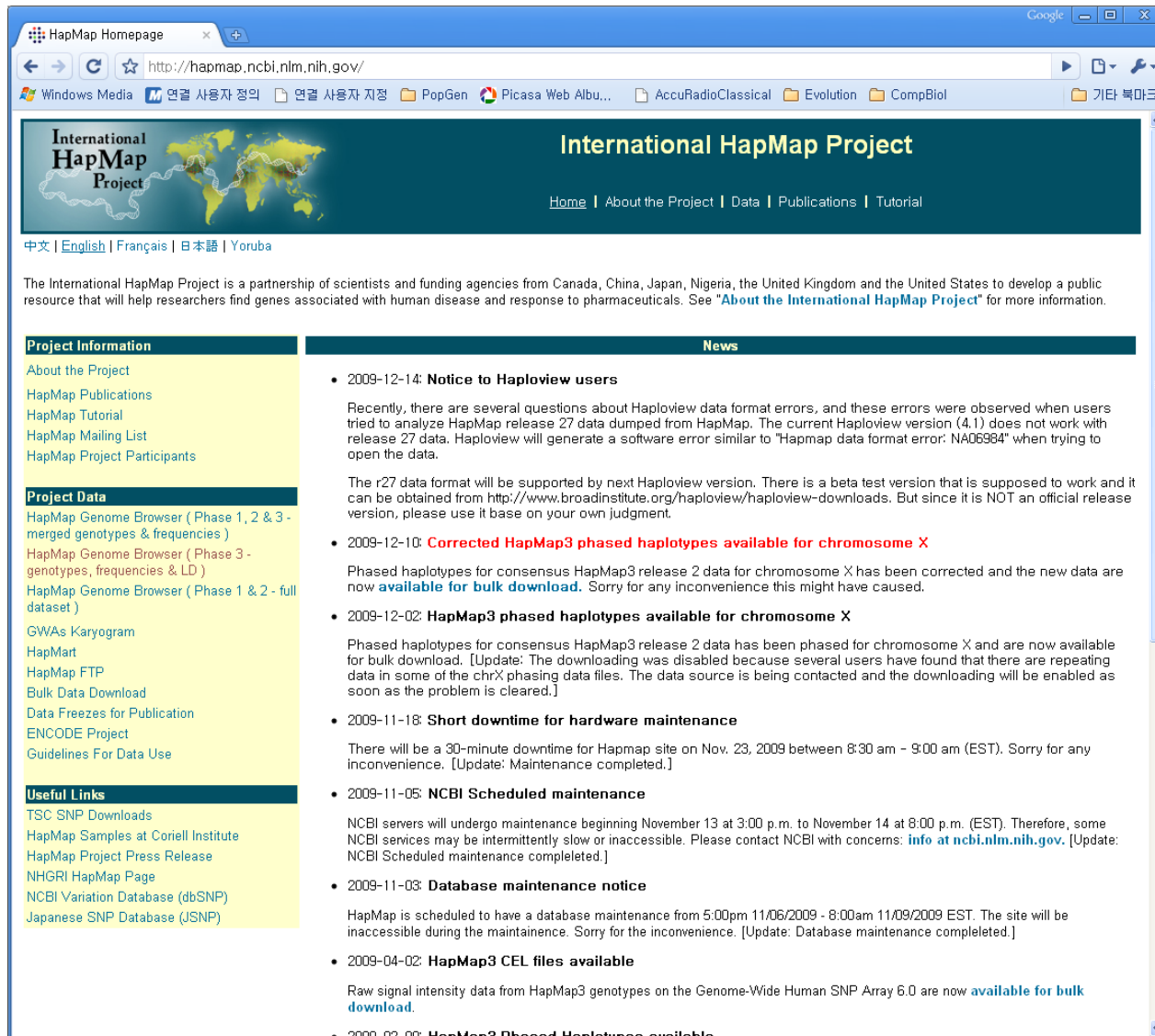
# Overview of GWAS



# Assumptions in GWAS

- Bi-allelic SNPs
- Common ancestors
- Linkage disequilibrium and haplotypes
- Common disease-common variant

# International HapMap Project



The screenshot shows the International HapMap Project homepage in a web browser. The browser's address bar displays the URL <http://hapmap.ncbi.nlm.nih.gov/>. The page features a dark blue header with the project logo on the left and the title "International HapMap Project" on the right. Below the title are navigation links: [Home](#), [About the Project](#), [Data](#), [Publications](#), and [Tutorial](#). A language selection bar includes [中文](#), [English](#), [Français](#), [日本語](#), and [Yoruba](#).

The main content area is divided into two columns. The left column contains a "Project Information" section with links for "About the Project", "HapMap Publications", "HapMap Tutorial", "HapMap Mailing List", and "HapMap Project Participants". Below this is a "Project Data" section with links for "HapMap Genome Browser (Phase 1, 2 & 3 - merged genotypes & frequencies)", "HapMap Genome Browser (Phase 3 - genotypes, frequencies & LD)", "HapMap Genome Browser (Phase 1 & 2 - full dataset)", "GWAs Karyogram", "HapMart", "HapMap FTP", "Bulk Data Download", "Data Freezes for Publication", "ENCODE Project", and "Guidelines For Data Use". A "Useful Links" section follows with links for "TSC SNP Downloads", "HapMap Samples at Coriell Institute", "HapMap Project Press Release", "NHGRI HapMap Page", "NCBI Variation Database (dbSNP)", and "Japanese SNP Database (JSNP)".

The right column features a "News" section with a list of updates:

- 2009-12-14: Notice to Haploview users**  
Recently, there are several questions about Haploview data format errors, and these errors were observed when users tried to analyze HapMap release 27 data dumped from HapMap. The current Haploview version (4.1) does not work with release 27 data. Haploview will generate a software error similar to "Hapmap data format error: NA06984" when trying to open the data.  
The r27 data format will be supported by next Haploview version. There is a beta test version that is supposed to work and it can be obtained from <http://www.broadinstitute.org/haploview/haploview-downloads>. But since it is NOT an official release version, please use it base on your own judgment.
- 2009-12-10: Corrected HapMap3 phased haplotypes available for chromosome X**  
Phased haplotypes for consensus HapMap3 release 2 data for chromosome X has been corrected and the new data are now **available for bulk download**. Sorry for any inconvenience this might have caused.
- 2009-12-02: HapMap3 phased haplotypes available for chromosome X**  
Phased haplotypes for consensus HapMap3 release 2 data has been phased for chromosome X and are now available for bulk download. [Update: The downloading was disabled because several users have found that there are repeating data in some of the chrX phasing data files. The data source is being contacted and the downloading will be enabled as soon as the problem is cleared.]
- 2009-11-18: Short downtime for hardware maintenance**  
There will be a 30-minute downtime for Hapmap site on Nov. 23, 2009 between 8:30 am - 9:00 am (EST). Sorry for any inconvenience. [Update: Maintenance completed.]
- 2009-11-05: NCBI Scheduled maintenance**  
NCBI servers will undergo maintenance beginning November 13 at 3:00 p.m. to November 14 at 8:00 p.m. (EST). Therefore, some NCBI services may be intermittently slow or inaccessible. Please contact NCBI with concerns: [info at ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov). [Update: NCBI Scheduled maintenance completed.]
- 2009-11-03: Database maintenance notice**  
HapMap is scheduled to have a database maintenance from 5:00pm 11/06/2009 - 8:00am 11/09/2009 EST. The site will be inaccessible during the maintenance. Sorry for the inconvenience. [Update: Database maintenance completed.]
- 2009-04-02: HapMap3 CEL files available**  
Raw signal intensity data from HapMap3 genotypes on the Genome-Wide Human SNP Array 6.0 are now **available for bulk download**.
- 2009-02-09: HapMap3 Phased Haplotypes available**

# Samples

- Matched case-control samples on age, sex, demographics
- Case: more severely affected individuals
- Control: low risk of disease, rather than population-based samples
- Common population structure
  - Population stratification

# Genotyping methods

- ~ 1 million SNPs on a chip
- Affymetrix or Illumina
  - Random equi-distant probes
  - Gene-dense probes
  - Haplotype tagging SNPs
  - Copy number probes

# Quality Control

- Poor markers
  - Violation of Hardy-Weinberg equilibrium
  - Genotype call rates  $< 95\%$
  - Minor allele frequency (MAF)  $< 0.01$
- Poor samples
  - Genotyping rate  $< 95\%$
  - Gender inconsistency
  - Cryptic relatedness

# Statistical tests

- Case-control
  - Allelic chisq test
  - Cochran-Armitage trend test
  - Logistic regression  
([http://www.well.ox.ac.uk/rmott/LECTURES/LOGISTIC\\_REGRESSION/Logistic%20Regression%20using%20R.ppt](http://www.well.ox.ac.uk/rmott/LECTURES/LOGISTIC_REGRESSION/Logistic%20Regression%20using%20R.ppt))
- Quantitative traits
  - Linear regression
- Covariate interations
  - Age, sex etc

# Case-control association test

## Chi square & OR

Genotype	aa	aA	AA	Total	aa	aA	AA	Total
Case	542	2062	2033	4637	0	292	4345	4637
Control	514	1905	1786	4205	0	381	3824	4205
Total	1056	3967	3819	8842	0	673	8169	8842

Allele	a	A	Total	a	A	Total
Case	3146	6128	9274	292	8982	9274
Control	2933	5477	8410	381	8029	8410
Total	6079	11605	17684	673	17011	17684

Odds (case)	$3146/6128=0.513$	$292/8982=0.0325$
Odds (control)	$2933/5477=0.5355$	$381/8029=0.04745$
Odds ratio	$0.513/0.5355=0.959$	$0.0325/0.04745=0.685$
P (ChiSQ)	0.183	1.619e-06

# Cochran-Armitage Trend Test

Genotype	aa	aA	AA	Sum
Cases	$r_0$	$r_1$	$r_2$	$R$
Contorls	$s_0$	$s_1$	$s_2$	$S$
Sum	$n_0$	$n_1$	$n_2$	$N$

aa	aA	AA	Sum
542	2062	2033	4637
514	1905	1786	4205
1056	3967	3819	8842

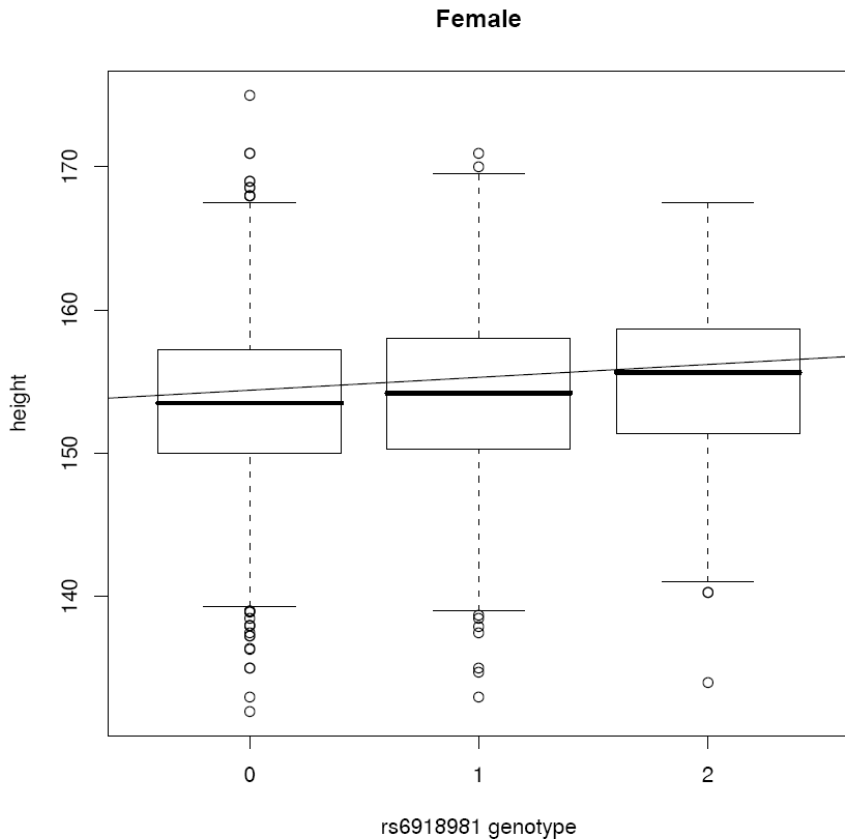
$$T := \sum_{i=0}^2 t_i(r_i S - s_i R),$$

$$\Pr(\text{Case}|\text{Genotype } i) = \Pr(\text{Case}|\text{Genotype } j) = n_i/N$$

- a dominant over A  
 $t = (1,1,0)$
- a recessive to A  
 $t = (0,1,1)$
- a and A additive  
 $t = (0,1,2)$

- Additive P = 0.1842
- Dominant P = 0.1941
- Recessive P = 0.4386

# Quantitative traits



- Genotypes coded (additive mode)
  - 0 major homozygotes
  - 1 heterozygotes
  - 2 minor homozygotes
- Linear regression
  - Intercept = 153.54
  - Slope = 0.6086
  - P value =  $2.05e-05$

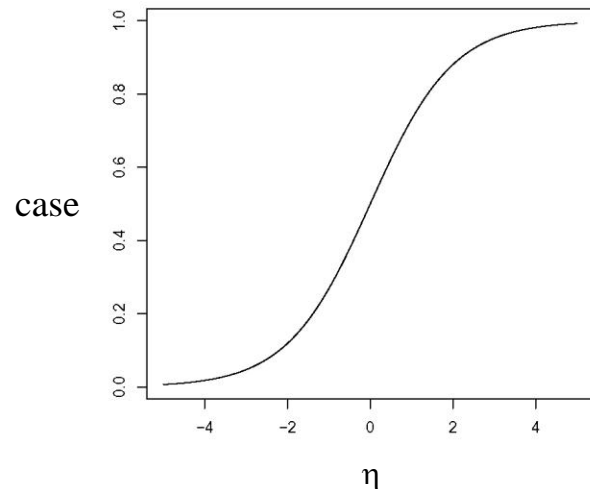
# Covariate adjustment

- Case-control
  - Logistic regression

$$\eta = \text{genotype} + \text{sex} + \text{age} + \varepsilon$$
$$\text{case} \sim \exp(\eta) / (1 + \exp(\eta))$$

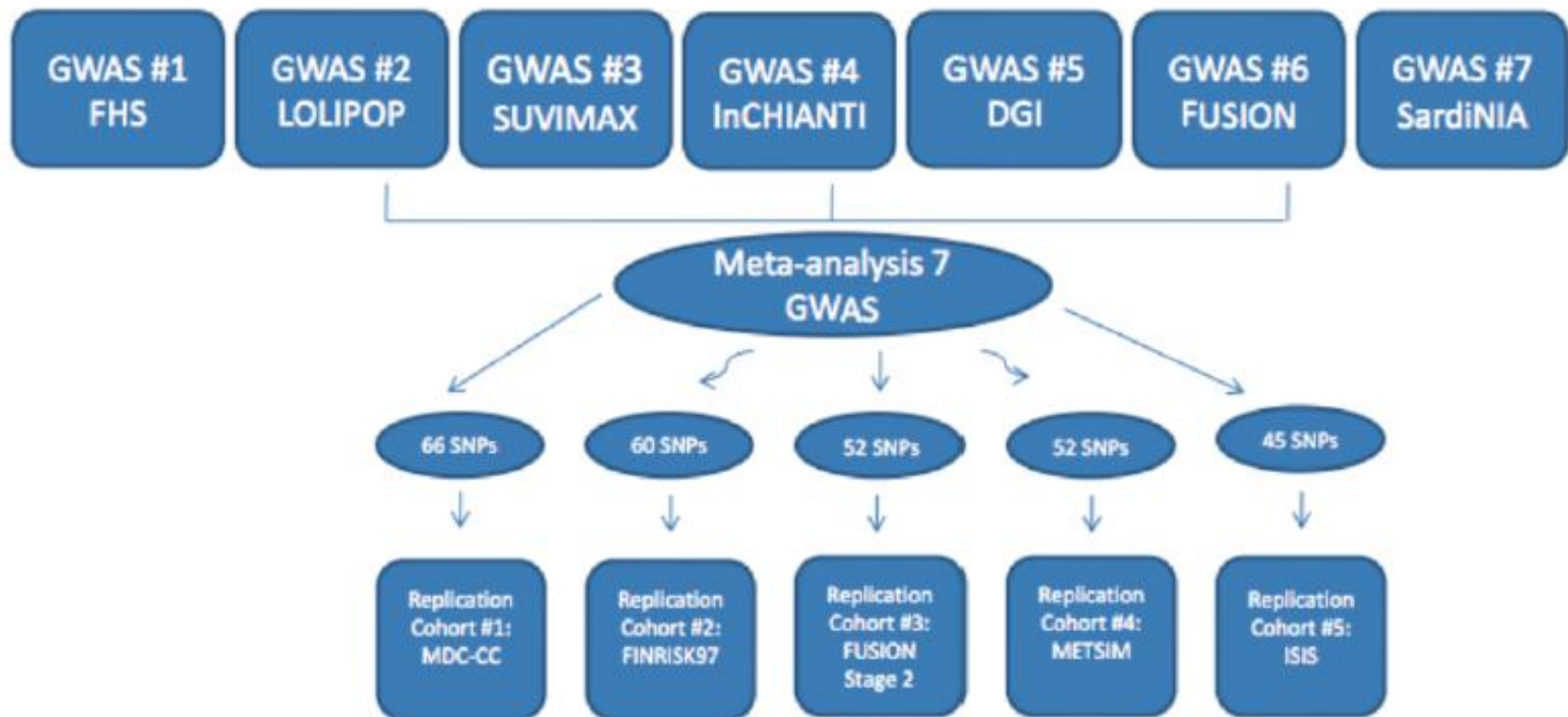
- Quantitative traits
  - Linear regression

$$\text{height} \sim \text{genotype} + \text{sex} + \text{age} + \varepsilon$$



# Imputation

- Genotypes not measured with SNP chips can be inferred by referencing HapMap haplotypes
- Increases marker density; helps define signal boundaries
- Facilitates merging datasets from different platforms; critical for meta analysis



GWA in ~19,840 individuals  
Follow-up in ~20,623 individuals

1. Introduction

2. Basic information

- [Citing PLINK](#)
- [Reporting problems](#)
- [What's new?](#)
- [PDF documentation](#)

3. Download and general notes

- [Stable download](#)
- [Development code](#)
- [General notes](#)
- [MS-DOS notes](#)
- [Unix/Linux notes](#)
- [Compilation](#)
- [Using the command line](#)
- [Viewing output files](#)
- [Version history](#)

4. Command reference table

- [List of options](#)
- [List of output files](#)
- [Under development](#)

5. Basic usage/data formats

- [Running PLINK](#)
- [PED files](#)
- [MAP files](#)
- [Transposed filesets](#)
- [Long-format filesets](#)
- [Binary PED files](#)
- [Alternate phenotypes](#)
- [Covariate files](#)
- [Cluster files](#)
- [Set files](#)

6. Data management

- [Recode](#)
- [Reorder](#)
- [Write SNP list](#)
- [Update SNP map](#)
- [Update allele information](#)
- [Force reference allele](#)
- [Update individuals](#)
- [Write covariate files](#)
- [Write cluster files](#)
- [Flip strand](#)
- [Scan for strand problem](#)
- [Merge two files](#)
- [Merge multiple files](#)
- [Extract SNPs](#)
- [Remove SNPs](#)
- [Zero out sets of genotypes](#)
- [Extract individuals](#)

**PLINK** is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

The focus of **PLINK** is purely on *analysis* of genotype/phenotype data, so there is no support for steps prior to this (e.g. study design and planning, generating genotype or CNV calls from raw data). Through integration with [gPLINK](#) and [Haploview](#), there is some support for the subsequent visualization, annotation and storage of results.

**PLINK** (one syllable) is being developed by Shaun Purcell at the Center for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH), and the [Broad Institute](#) of Harvard & MIT, with the support of others.

**New in 1.07:** meta-analysis, result annotation and analysis of dosage data.

Quick links

[PLINK tutorial](#)

[gPLINK](#)

[Join e-mail list](#)

[Resources](#)

[FAQs](#) | [PDF](#)

[Citing PLINK](#)

[Bugs, questions?](#)

Data management

- [Read data in a variety of formats](#)
- [Recode and reorder files](#)
- [Merge two or more files](#)
- [Extracts subsets \(SNPs or individuals\)](#)
- [Flip strand of SNPs](#)
- [Compress data in a binary file format](#)

Summary statistics for quality control

- [Allele, genotypes frequencies, HWE tests](#)
- [Missing genotype rates](#)
- [Inbreeding, IBS and IBD statistics for individuals and pairs of individuals](#)
- [non-Mendelian transmission in family data](#)
- [Sex checks based on X chromosome SNPs](#)
- [Tests of non-random genotyping failure](#)

Population stratification detection

- [Complete linkage hierarchical clustering](#)
- [Handles virtually unlimited numbers of SNPs](#)
- [Multidimensional scaling analysis to visualise substructure](#)
- [Significance test for whether two individuals belong to the same population](#)
- [Constrain cluster solution by phenotype, cluster size and/or external matching criteria](#)
- [Perform subsequent association analyses conditional on cluster solution](#)

# An example

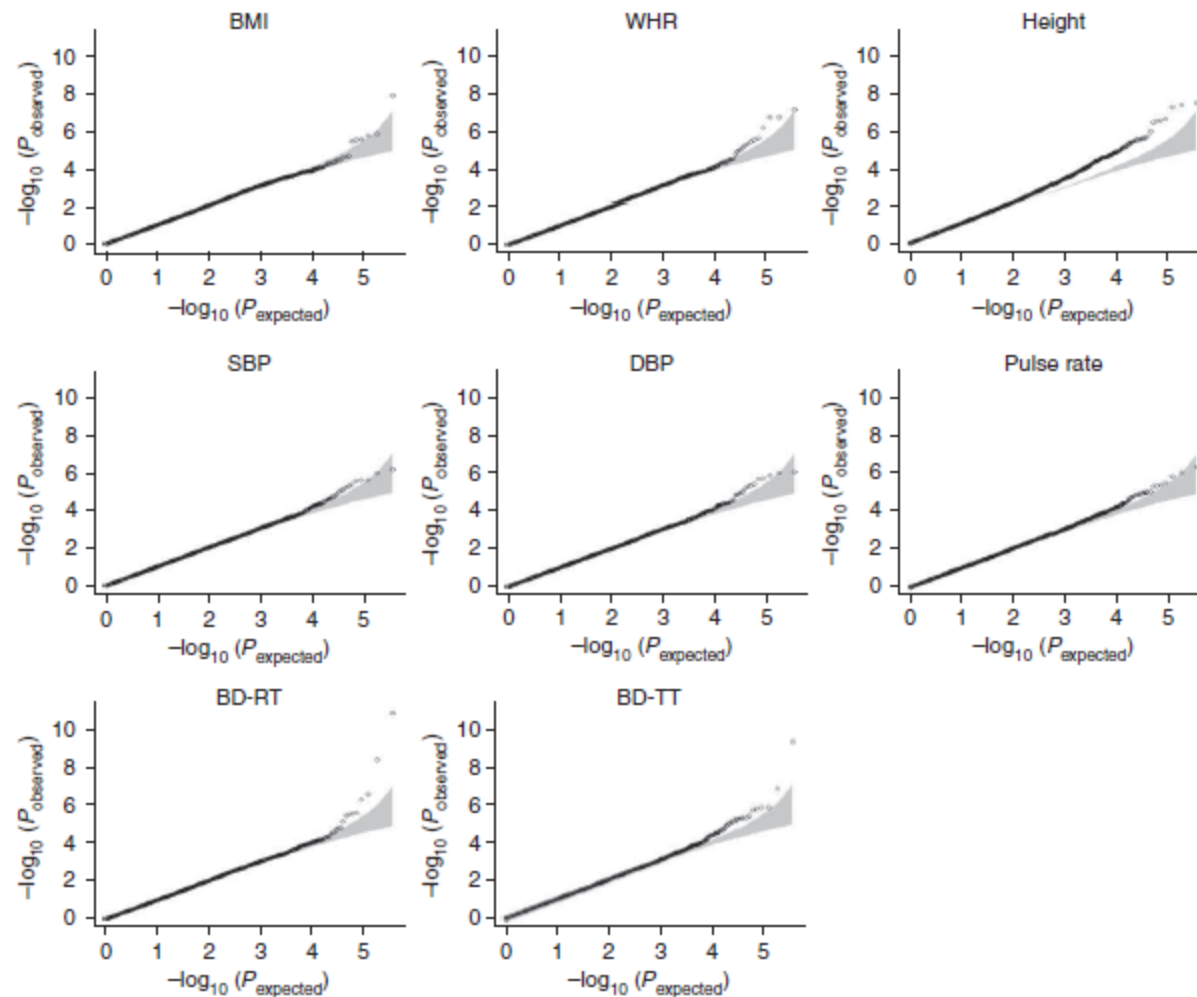
## A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits

Yoon Shin Cho<sup>1</sup>, Min Jin Go<sup>1</sup>, Young Jin Kim<sup>1</sup>, Jee Yeon Heo<sup>1</sup>, Ji Hee Oh<sup>1</sup>, Hyo-Jeong Ban<sup>1</sup>, Dankyu Yoon<sup>2</sup>, Mi Hee Lee<sup>1</sup>, Dong-Joon Kim<sup>1</sup>, Miey Park<sup>1</sup>, Seung-Hun Cha<sup>1</sup>, Jun-Woo Kim<sup>1</sup>, Bok-Ghee Han<sup>1</sup>, Haesook Min<sup>1</sup>, Younjhin Ahn<sup>1</sup>, Man Suk Park<sup>1</sup>, Hye Ree Han<sup>1</sup>, Hye-Yoon Jang<sup>3</sup>, Eun Young Cho<sup>3</sup>, Jong-Eun Lee<sup>3</sup>, Nam H Cho<sup>4</sup>, Chol Shin<sup>5</sup>, Taesung Park<sup>2,6</sup>, Ji Wan Park<sup>7</sup>, Jong-Keuk Lee<sup>8</sup>, Lon Cardon<sup>9</sup>, Geraldine Clarke<sup>10</sup>, Mark I McCarthy<sup>10,11</sup>, Jong-Young Lee<sup>1</sup>, Jong-Koo Lee<sup>12</sup>, Bermseok Oh<sup>1,13</sup> & Hyung-Lae Kim<sup>1</sup>

To identify genetic factors influencing quantitative traits of biomedical importance, we conducted a genome-wide association study in 8,842 samples from population-based cohorts recruited in Korea. For height and body mass index, most variants detected overlapped those reported in European samples. For the other traits examined, replication of promising GWAS signals in 7,861 independent Korean samples identified six previously unknown loci. For pulse rate, signals reaching genome-wide significance mapped to chromosomes 1q32 (rs12731740,  $P = 2.9 \times 10^{-9}$ ) and 6q22 (rs12110693,  $P = 1.6 \times 10^{-9}$ ), with the latter  $\sim 400$  kb from the coding sequence of *GJA1*. For systolic blood pressure, the most compelling association involved chromosome 12q21 and variants near the *ATP2B1* gene (rs17249754,  $P = 1.3 \times 10^{-7}$ ). For waist-hip ratio, variants on chromosome 12q24 (rs2074356,  $P = 7.8 \times 10^{-12}$ ) showed convincing associations, although no regional transcript has strong biological candidacy. Finally, we identified two loci influencing bone mineral density at multiple sites. On chromosome 7q31, rs7776725 (within the *FAM3C* gene) was associated with bone density at the radius ( $P = 1.0 \times 10^{-11}$ ), tibia ( $P = 1.6 \times 10^{-6}$ ) and heel ( $P = 1.9 \times 10^{-10}$ ). On chromosome 7p14, rs1721400 (mapping close to *SFRP4*, a frizzled protein gene) showed consistent associations at the same three sites ( $P = 2.2 \times 10^{-3}$ ,  $P = 1.4 \times 10^{-7}$  and  $P = 6.0 \times 10^{-4}$ , respectively). This large-scale GWA analysis of well-characterized Korean population-based samples highlights previously unknown biological pathways.

# Korea Association Resource (KARE) project

- Affymetrix 5.0 genotypes on 10,004 individuals (ages 40~69)
- 352,228 SNPs passed QC
  - 38,364 markers violated HWE ( $P < 10^{-6}$ )
  - 17,926 genotype call rates  $< 95\%$
  - 92,050 MAF  $< 0.01$
- 8,842 individuals passed QC
  - 11 sample contamination
  - 41 gender inconsistency
  - 608 cryptic relatedness
  - 101 serious concomitant illness

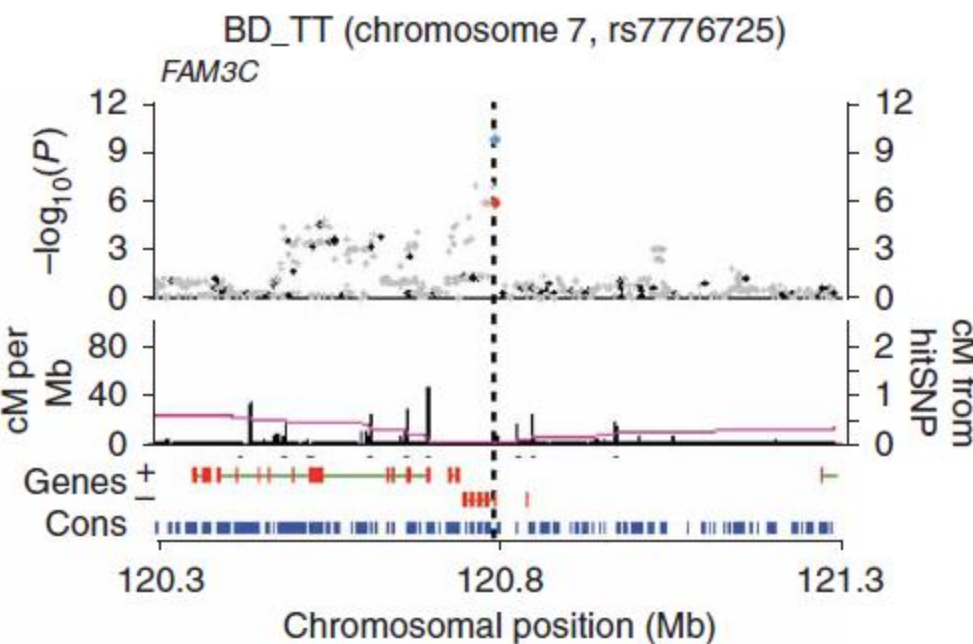
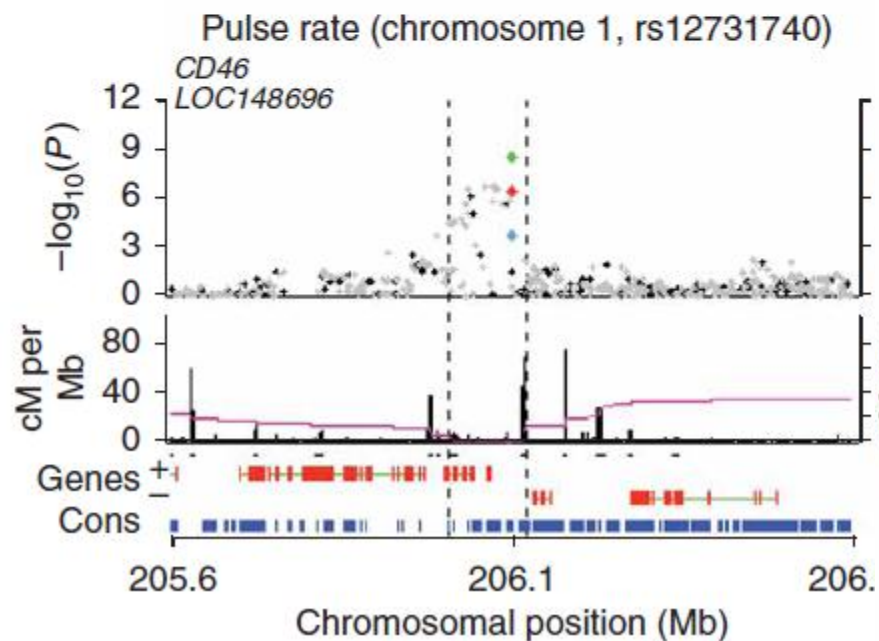
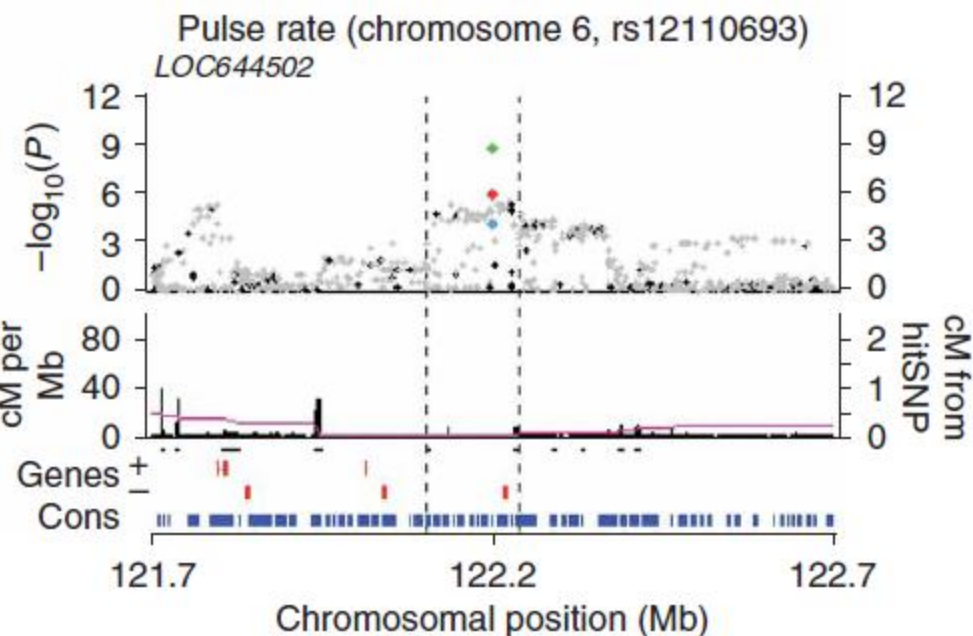
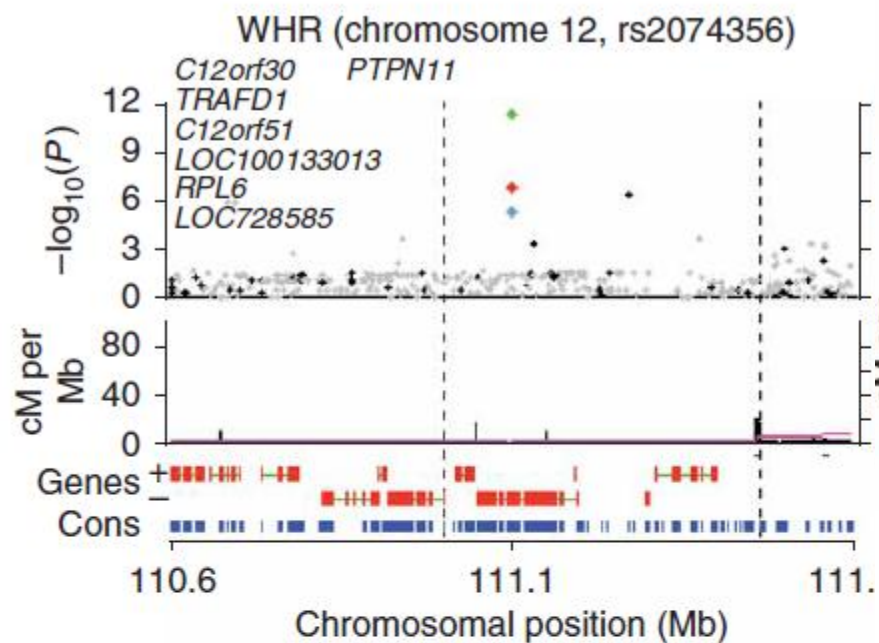


**Figure 2** Quantile-quantile plots for the eight quantitative traits. The observed  $P$  values (y axis) were compared with the expected  $P$  values under the null distribution (x axis) for each trait. The shaded region represents the 95% concentration band. (BMI, body mass index; WHR, waist hip ratio; SBP, systolic blood pressure; DBP, diastolic blood pressure; BD-RT, bone density estimated by T-score at distal radius; BD-TT, bone density estimated by T-score at midshaft tibia).

# GWAS, replication, literature

**Table 1** Association between SNPs and quantitative traits in the population with GWA data and in the replication population

Trait	RS ID	Class	Locus	Nearby genes <sup>a</sup>	Minor allele	KARE GWAS			Sample size for 80% power <sup>b</sup>	Trend <i>P</i> value			Effect size of combined (β ± s.e.m.)	Meta-analysis heterogeneity Q (P)	Previously published
						MAF	Effect size (β ± s.e.m.)	Variation/copy		GWAS (n = 8,842)	Replication (n = 7,861)	Combined <sup>c</sup> (n = 16,703)			
BMI	rs17178527	Unknown	6q24.1	<i>LOC729076</i>	A	0.25	-0.312 ± 0.055	-0.310	2,127	1.2E-08	3.9E-01	3.7E-04	-0.142 ± 0.040	20.63 (5.6E-06)	
	rs9939609	Intron	16q12.2	<i>FTO</i>	A	0.13	0.335 ± 0.07	0.200	3,190	1.7E-06	2.2E-02	1.5E-07	0.235 ± 0.045	5.73 (1.7E-02)	Frayling <i>et al.</i> <sup>7</sup>
WHR	<b>rs2074356</b>	Intron	12q24.13	<i>C12orf51</i>	T	0.15	-0.007 ± 0.001	-0.005	1,757	1.8E-07	7.6E-06	7.8E-12	-0.006 ± 0.001	0.09 (7.7E-01)	
Height	rs17089410	Unknown	13q21.33		T	0.14	0.006 ± 0.001	0.005	1,852	6.1E-06	5.3E-01	4.4E-03	0.003 ± 0.001	12.59 (3.9E-04)	
	rs6918981	Unknown	6p21.31	<i>HMGAI</i>	G	0.21	0.542 ± 0.098	1.030	6,113	3.2E-08	2.5E-02	3.3E-08	0.401 ± 0.072	4.25 (3.9E-02)	Gudbjartsson <i>et al.</i> (rs1776897) <sup>23</sup>
	rs17038182	Unknown	1p12		C	0.42	-0.451 ± 0.082	-0.320	5,977	4.3E-08	1.2E-01	4.7E-07	-0.303 ± 0.060	6.75 (9.4E-03)	Weedon <i>et al.</i> (rs6440003) <sup>4</sup>
	rs10513137	Intron	3q23	<i>ZBTB38</i>	A	0.26	0.492 ± 0.091	0.760	6,363	5.6E-08	1.8E-05	5.6E-12	0.461 ± 0.067	0.24 (6.2E-01)	
	rs13273123	Intron	8q12.1	<i>PLAG1</i>	G	0.07	-0.781 ± 0.16	-0.165	7,858	1.1E-06	2.0E-04	1.0E-09	-0.710 ± 0.116	0.40 (5.2E-01)	Gudbjartsson <i>et al.</i> (rs10958476) <sup>23</sup>
	rs600130	Intron	9q22.32	<i>FBP2</i>	G	0.15	-0.529 ± 0.113	-0.330	8,367	2.7E-06	5.1E-01	9.9E-05	-0.316 ± 0.081	7.75 (5.4E-03)	
	rs2079795	Unknown	17q23.2	<i>BCAS3, TBX2</i>	A	0.33	0.399 ± 0.085	0.315	8,357	2.9E-06	-	-	-	-	Gudbjartsson <i>et al.</i> (rs757608) <sup>23</sup>
	rs3791675	Intron	2p16.1	<i>EFEMP1</i>	G	0.22	0.445 ± 0.096	0.310	8,346	3.6E-06	1.0E-04	1.7E-09	0.424 ± 0.070	0.11 (7.4E-01)	Weedon <i>et al.</i> <sup>4</sup>
	rs41464348	Intron	2p22.3	<i>LTBP1</i>	T	0.35	-0.370 ± 0.082	-0.195	9,414	7.4E-06	-	-	-	-	
SBP	<b>rs17249754</b>	Unknown	12q21.33	<i>ATP2B1</i>	A	0.37	-1.309 ± 0.266	-1.260	2,904	9.1E-07	9.9E-03	1.3E-07	-1.064 ± 0.201	1.65 (2.0E-01)	
	rs715987	Unknown	10p15.1		C	0.15	-1.660 ± 0.362	-1.695	3,337	4.5E-06	4.8E-01	6.4E-03	-0.741 ± 0.272	12.77 (3.5E-04)	
DBP	rs17249754	Unknown	12q21.33	<i>ATP2B1</i>	A	0.37	-0.882 ± 0.181	-0.860	2,748	1.2E-06	8.6E-02	3.0E-06	-0.630 ± 0.135	3.82 (5.1E-02)	
Pulse rate	<b>rs12731740</b>	Unknown	1q32.2	<i>CD46, LOC148696</i>	T	0.10	0.993 ± 0.195	1.030	2,448	3.7E-07	2.0E-04	2.9E-09	1.085 ± 0.183	0.27 (6.0E-01)	
	rs12110693	Unknown	6q22.31	<i>LOC644502</i>	A	0.49	0.573 ± 0.118	0.620	2,729	1.3E-06	7.0E-05	1.6E-09	0.661 ± 0.109	0.69 (4.1E-01)	
	rs11576175	Intron	1q21.2	<i>CTSS</i>	A	0.24	0.630 ± 0.141	0.855	3,098	8.3E-06	5.8E-02	3.7E-05	0.534 ± 0.129	0.59 (4.4E-01)	
BD-RT	<b>rs7776725</b>	Intron	7q31.31	<i>FAM3C</i>	C	0.13	0.222 ± 0.033	0.212	1,609	1.0E-11	1.9E-10 <sup>d</sup>	-	-	-	
	rs9525667	Unknown	13q14.11		T	0.43	-0.103 ± 0.022	-0.110	3,549	3.1E-06	3.5E-01 <sup>d</sup>	-	-	-	
BD-TT	<b>rs7776725</b>	Intron	7q31.31	<i>FAM3C</i>	C	0.13	0.155 ± 0.032	0.201	1,609	1.6E-06	1.9E-10 <sup>d</sup>	-	-	-	
	<b>rs1721400</b>	Unknown	7p14.1	<i>TXNDC3, SFRP4, EPDR1</i>	T	0.17	-0.149 ± 0.028	-0.136	2,987	1.4E-07	6.0E-04 <sup>d</sup>	-	-	-	
	rs550677	NearGene-5	12q24.31	<i>TMEM132B</i>	T	0.17	0.13 ± 0.028	0.195	3,979	5.0E-06	1.4E-01 <sup>d</sup>	-	-	-	
	rs6974574	Unknown	7p14.1		A	0.30	-0.105 ± 0.023	-0.110	3,774	7.9E-06	-	-	-	-	



# NHGRI GWAS Catalog

Genome.gov | A Catal... x

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

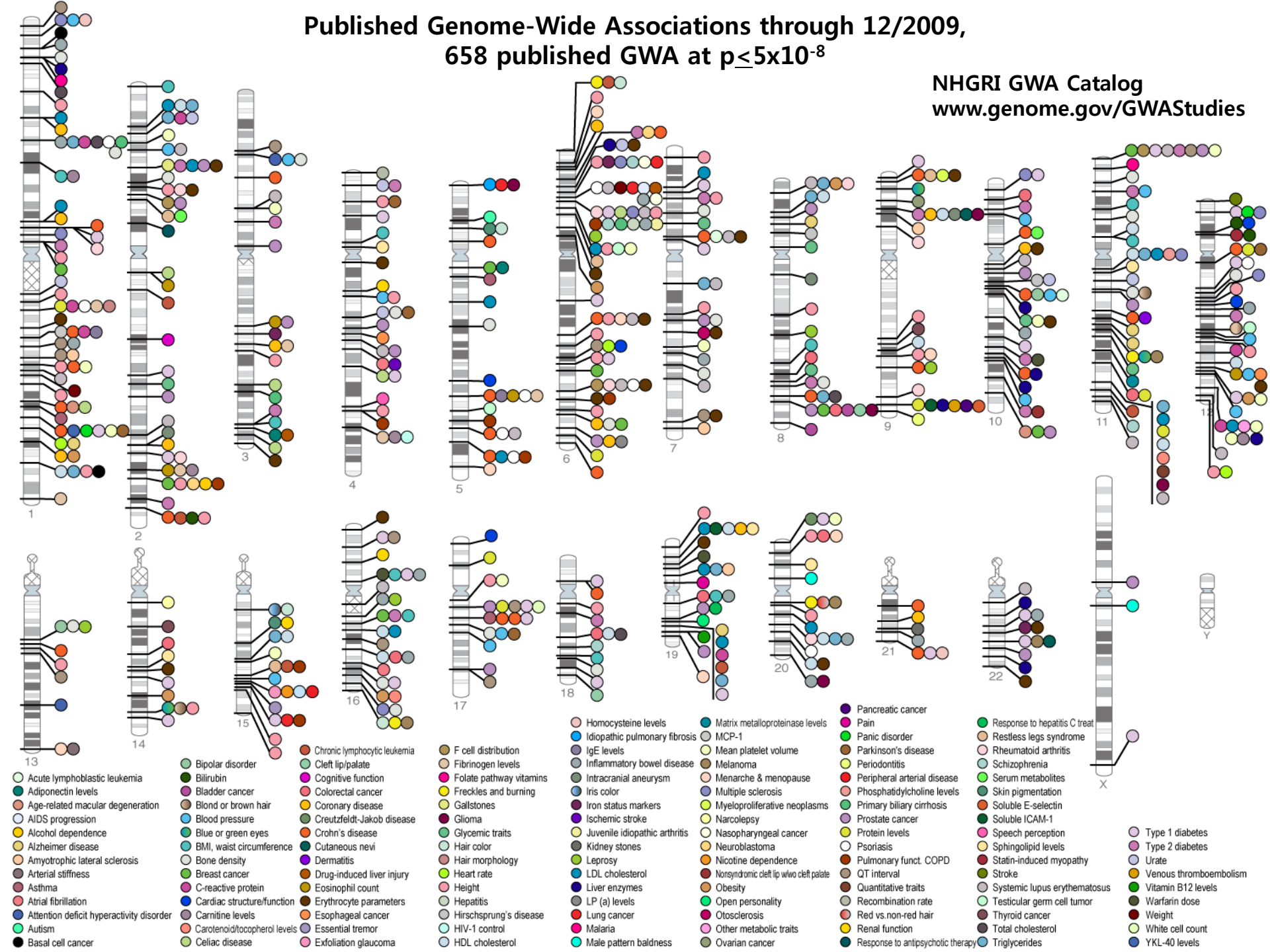
Windows Media 연결 사용자 정의 연결 사용자 지정 PopGen Picasa Web Album AccuRadioClassical Evolution CompBiol 기타 북마크

As of 04/02/10, this table includes 533 publications and 2540 SNPs.

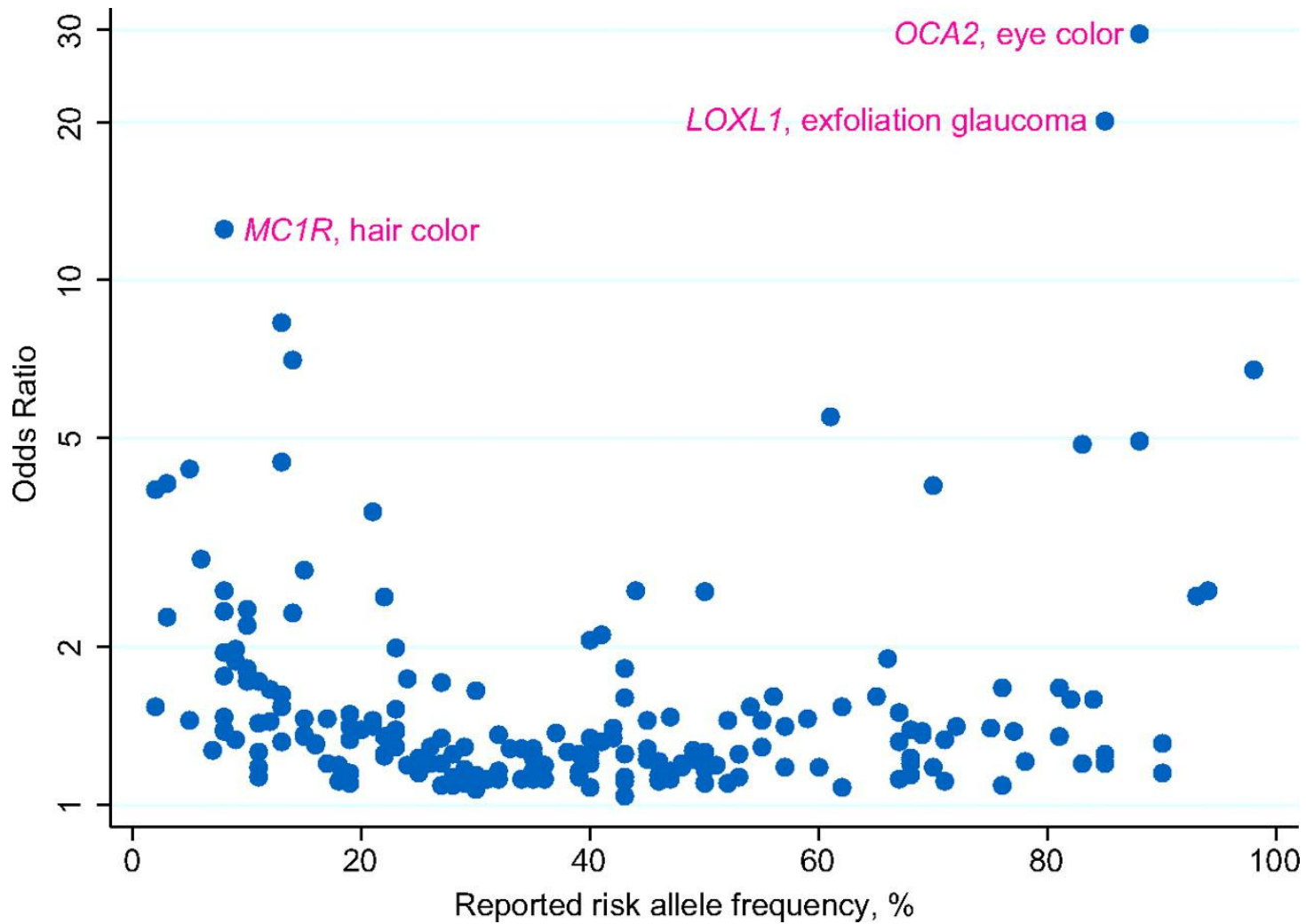
Date Added to Catalog (since 11/25/08)	First Author/Date/ Journal/Study	Disease/Trait	Initial Sample Size	Replication Sample Size	Region	Reported Gene(s)	Strongest SNP-Risk Allele	Risk Allele Frequency in Controls	P-value	OR or beta-coefficient and [95% CI]	Platform [SNPs passing QC]	CNV
03/29/10	Li March 19, 2010 <i>Lancet Oncol</i> <a href="#">Genetic variants and risk of lung cancer in never smokers: a genome-wide association study</a>	Lung cancer	377 cases, 377 matched controls	511 cases, 1,007 controls	13q31.3	<i>GPCS</i>	<a href="#">rs2352028-A</a>	0.26	$6 \times 10^{-6}$	1.46 [1.26-1.70]	Illumina [331,918]	N
04/02/10	Medland March 18, 2010 <i>Am J Hum Genet</i> <a href="#">A Variant in LIN28B Is Associated with 2D:4D Finger-Length Ratio, a Putative Retrospective Biomarker of Prenatal Testosterone Exposure</a>	Digit length ratio	2,889 European children and adolescents	3,659 European children	6q16.3	<i>LIN28B</i>	<a href="#">rs314277-A</a>	0.15	$2 \times 10^{-6}$	.63 [0.41-0.85] increase in mean 2D:4D	Illumina [310,613]	N
04/02/10	Nakajima March 18, 2010 <i>PLoS ONE</i> <a href="#">New Sequence Variants in HLA Class II/III Region Associated with Susceptibility to Knee Osteoarthritis Identified by Genome-Wide Association Study</a>	Knee osteoarthritis	899 Japanese cases, 3,396 Japanese controls	167 Japanese cases, 347 Japanese controls, 243 Spanish cases, 426 Spanish controls, 570 Greek cases, 645 Greek controls	6p21.32	<i>BTNL2, HLA-DQA2, HLA-DQB1</i>	<a href="#">rs10947262-I</a>	0.42	$5 \times 10^{-9}$	1.31 [1.20-1.44]	Illumina [459,393]	N
03/26/10	Smith March 15, 2010 <i>Circulation</i> <a href="#">Novel Associations of Multiple Genetic Loci With Plasma Levels of Factor VII, Factor VIII, and von Willebrand Factor. The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium</a>	Plasma coagulation factors	Up to 23,608 European ancestry individuals	Up to 7,604 European ancestry individuals	20q11.22 6q24.3	<i>PROCR</i> <i>STXBPS</i>	<a href="#">rs867186-G</a> <a href="#">rs9390459-A</a>	0.101 0.442	$6 \times 10^{-37}$ (FVII) $1 \times 10^{-22}$ (vWF)	NR 4.8 [2.1-7.5] % decrease	Affymetrix & Illumina [~2.6 million] (imputed)	N
03/24/10	Franke	Ulcerative colitis	1,043 German	2,539 European	1p36.13	<i>OUT3</i>	<a href="#">rs4654925-</a>	0.52	$6 \times 10^{-22}$	1.41 [1.30-1.54]	Affymetrix	N

# Published Genome-Wide Associations through 12/2009, 658 published GWA at $p \leq 5 \times 10^{-8}$

NHGRI GWA Catalog  
[www.genome.gov/GWASudies](http://www.genome.gov/GWASudies)

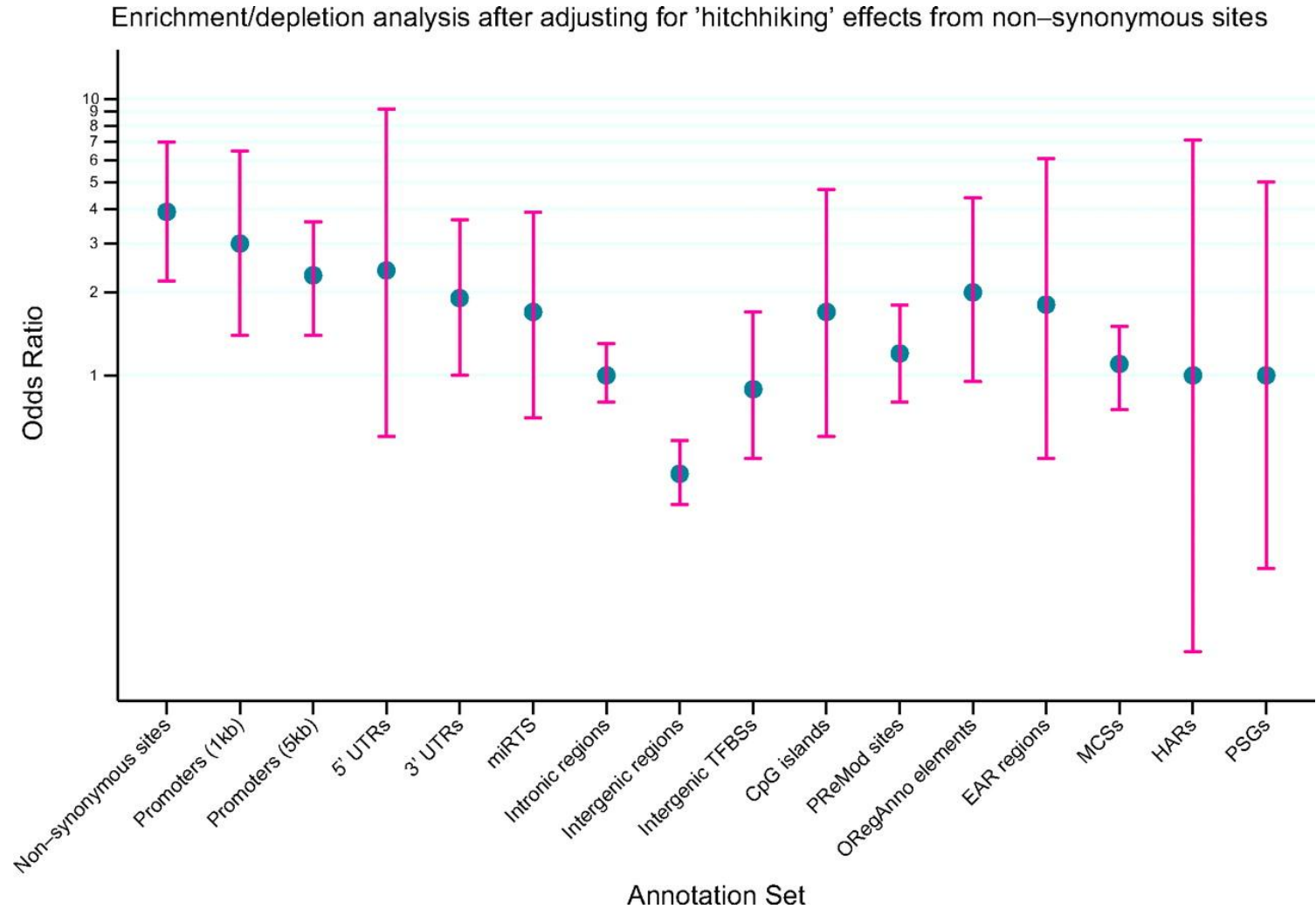


# Published odds ratios for discrete traits by reported risk allele frequencies.



Hindorff L A et al. PNAS 2009;106:9362-9367

# Odds ratios for TAS block enrichment/depletion analysis after adjusting for “hitchhiking” effects from nonsynonymous sites.



Hindorff L A et al. PNAS 2009;106:9362-9367

# Hundreds of GWAS applications tells us

- Many common variants of highly significant disease association have been found
- They confer relatively small increments in risk (1.0~1.5 fold)
- They explain only a small portion of heritability
  - Human height is estimated to have 80% heritability
  - About 5% of phenotype variance is explained based on  $>10^4$  people

# Excuses for the missing heritability

- Large numbers of variants of smaller effect yet to be found
- Rarer variants (possibly with larger effect)
- Structural variants poorly captured by existing arrays
- Low power to detect gene-gene interactions
- Inadequate accounting for shared environment among relatives

# Highly significant signals are found, but difficult to discuss biology

Trait	RS ID	Class	Locus	Nearby genes <sup>a</sup>	Minor allele	MAF	GWAS ( <i>n</i> = 8,842)	
BMI	rs17178527	Unknown	6q24.1	<i>LOC729076</i>	A	0.25	1.2E-08	
	rs9939609	Intron	16q12.2	<i>FTO</i>	A	0.13	1.7E-06	
WHR	<b>rs2074356</b>	Intron	12q24.13	<i>C12orf51</i>	T	0.15	1.8E-07	
	rs17089410	Unknown	13q21.33		T	0.14	6.1E-06	
Height	rs6918981	Unknown	6p21.31	<i>HMGA1</i>	G	0.21	3.2E-08	
	rs17038182	Unknown	1p12		C	0.42	4.3E-08	
	rs10513137	Intron	3q23	<i>ZBTB38</i>	A	0.26	5.6E-08	
	rs13273123	Intron	8q12.1	<i>PLAG1</i>	G	0.07	1.1E-06	
	rs600130	Intron	9q22.32	<i>FBP2</i>	G	0.15	2.7E-06	
	rs2079795	Unknown	17q23.2	<i>BCAS3, TBX2</i>	A	0.33	2.9E-06	
	rs3791675	Intron	2p16.1	<i>EFEMP1</i>	G	0.22	3.6E-06	
	rs41464348	Intron	2p22.3	<i>LTBP1</i>	T	0.35	7.4E-06	
	SBP	<b>rs17249754</b>	Unknown	12q21.33	<i>ATP2B1</i>	A	0.37	9.1E-07
		rs715987	Unknown	10p15.1		C	0.15	4.5E-06
DBP	rs17249754	Unknown	12q21.33	<i>ATP2B1</i>	A	0.37	1.2E-06	
Pulse rate	<b>rs12731740</b>	Unknown	1q32.2	<i>CD46, LOC148696</i>	T	0.10	3.7E-07	
	<b>rs12110693</b>	Unknown	6q22.31	<i>LOC644502</i>	A	0.49	1.3E-06	
	rs11576175	Intron	1q21.2	<i>CTSS</i>	A	0.24	8.3E-06	
BD-RT	<b>rs7776725</b>	Intron	7q31.31	<i>FAM3C</i>	C	0.13	1.0E-11	
	rs9525667	Unknown	13q14.11		T	0.43	3.1E-06	
BD-TT	<b>rs7776725</b>	Intron	7q31.31	<i>FAM3C</i>	C	0.13	1.6E-06	
	<b>rs1721400</b>	Unknown	7p14.1	<i>TXNDC3, SFRP4, EPDR1</i>	T	0.17	1.4E-07	
	rs550677	NearGene-5	12q24.31	<i>TMEM132B</i>	T	0.17	5.0E-06	
	rs6974574	Unknown	7p14.1		A	0.30	7.9E-06	

# Rare Variants Create Synthetic Genome-Wide Associations

Samuel P. Dickson<sup>1,2</sup>, Kai Wang<sup>3</sup>, Ian Krantz<sup>3,4,5</sup>, Hakon Hakonarson<sup>3,4,5</sup>, David B. Goldstein<sup>1\*</sup>

**1** Institute for Genome Sciences and Policy, Center for Human Genome Variation, Duke University, Durham, North Carolina, United States of America, **2** Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America, **3** Center for Applied Genomics, Children's Hospital of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **4** Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, United States of America, **5** Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America

## Abstract

Genome-wide association studies (GWAS) have now identified at least 2,000 common variants that appear associated with common diseases or related traits (<http://www.genome.gov/gwastudies>), hundreds of which have been convincingly replicated. It is generally thought that the associated markers reflect the effect of a nearby common (minor allele frequency >0.05) causal site, which is associated with the marker, leading to extensive resequencing efforts to find causal sites. We propose as an alternative explanation that variants much less common than the associated one may create "synthetic associations" by occurring, stochastically, more often in association with one of the alleles at the common site versus the other allele. Although synthetic associations are an obvious theoretical possibility, they have never been systematically explored as a possible explanation for GWAS findings. Here, we use simple computer simulations to show the conditions under which such synthetic associations will arise and how they may be recognized. We show that they are not only possible, but inevitable, and that under simple but reasonable genetic models, they are likely to account for or contribute to many of the recently identified signals reported in genome-wide association studies. We also illustrate the behavior of synthetic associations in real datasets by showing that rare causal mutations responsible for both hearing loss and sickle cell anemia create genome-wide significant synthetic associations, in the latter case extending over a 2.5-Mb interval encompassing scores of "blocks" of associated variants. In conclusion, uncommon or rare genetic variants can easily create synthetic associations that are credited to common variants, and this possibility requires careful consideration in the interpretation and follow up of GWAS signals.

We propose Gene-Set based approach

- Test gene-sets such as Gene Ontology biological processes, molecular signatures, etc
  - Designed to be biology-friendly
- Set-wise tests may be robust to different population structures
  - Weak but consistent associations may be detected

# Program submitted to NAR Webserver issue & under review

## GSA-SNP

Nucleic Acids Research

### Download

#### Program

Requirement: JRE (Java runtime environment) 1.6.0 or greater

- [GSA-SNP program](#) (stable version, about 120 MB)
- [GSA-SNP program](#) (development version)
- [manual](#)
- [Supplementary material](#)

#### Examples

- SNP
  - [100 permutations](#) (about 145 MB)
  - [without permutation](#)
- gene
  - [100 permutations](#)
  - [without permutation](#)
- haplotype
  - [without permutation](#)
- all
  - [download all above examples](#) (about 155 MB)

### Contact

E-mail to: [Dr. Dougu Nam](#) or [Dr. Sangsoo Kim](#)

Updated: 2010/02/06

GSA-SNP: a general approach for gene set analysis of polymorphisms

Dougu Nam<sup>1,+</sup>, Jin Kim<sup>2,+</sup>, Seon-Young Kim<sup>3</sup>, Sangsoo Kim<sup>4,\*</sup>

<sup>1</sup>*Division of Computational Mathematics, National Institute for Mathematical Sciences*

<sup>2</sup>*School of Computer Science and Engineering, Seoul National University*

<sup>3</sup>*Medical Genomics Research Center, Korea Research Institute for Bioscience and Biotechnology*

<sup>4</sup>*Department of Bioinformatics and Life Science, Soongsil University*

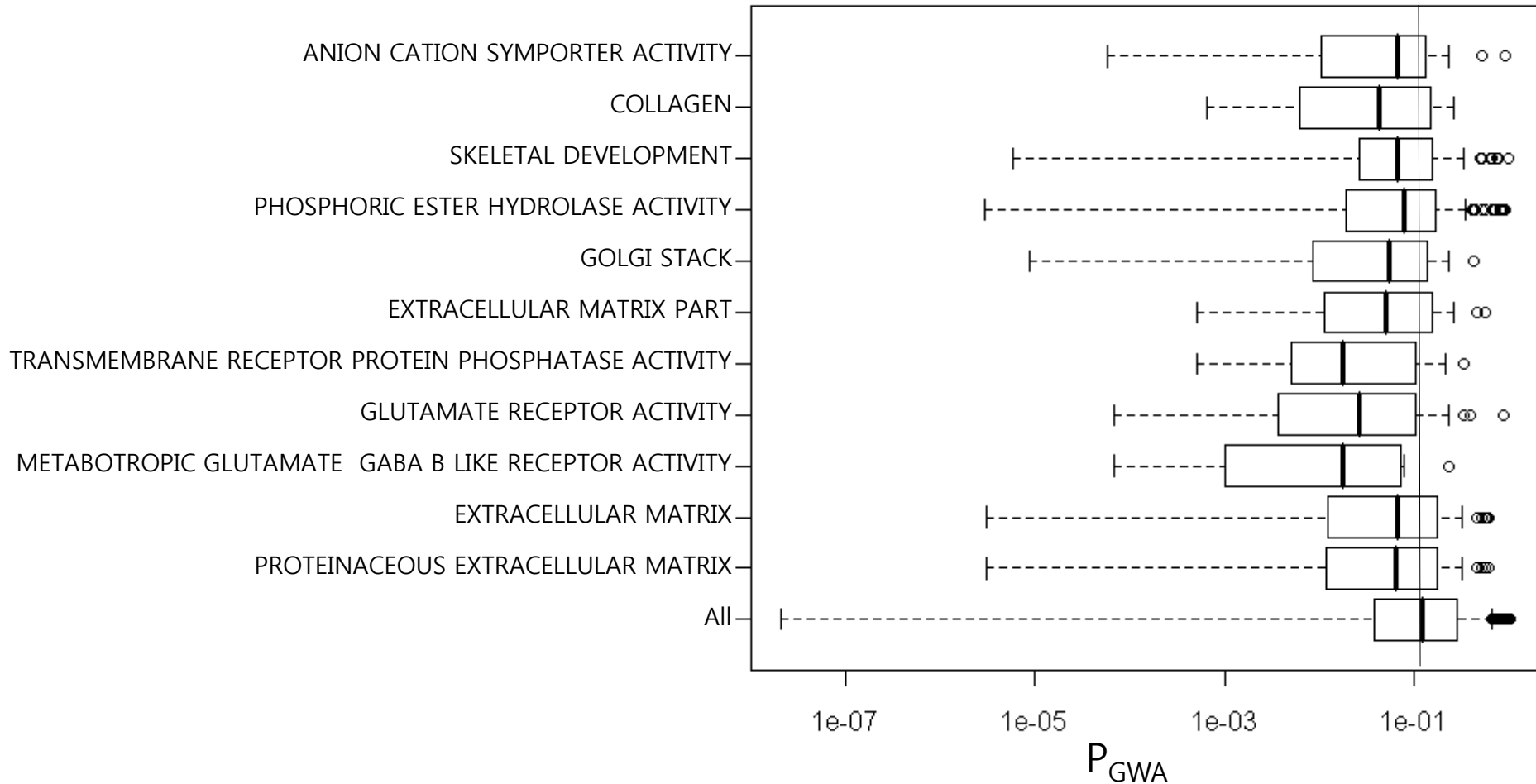
<sup>+</sup>These authors contributed equally to this work

<sup>\*</sup>Corresponding author

# Procedure

- PLINK association runs
  - The KARE genotypes were supplemented by imputing SNP genotypes based on those of the JPT+CHB panel of the HapMap Phase II
  - Additive linear regression model adjusted by age+sex
- SNPs within a gene boundary were summarized (2<sup>nd</sup> best; 20kb cushion)
- Another summarization by GO
  - Multiple testing corrected P-values are reported

# Moderate but consistent associations were detected in some Gene Ontology sets



# Literature survey

PROTEINACEOUS EXTRACELLULAR MATRIX	"A key biological function in height regulation" by Weedon et al. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet, 40, 575-583.
EXTRACELLULAR MATRIX	
METABOTROPIC GLUTAMATE GABA B LIKE RECEPTOR ACTIVITY	GRIA1, one of the members, was implicated near a loci associated with height in Croatian population. Endogenous activation of metabotropic glutamate receptors is known to modulate GABAergic transmission of gonadotropin-releasing hormone (GnRH) neurons. Moreover, treatment with a GnRH agonist in short adolescents increased adult height
GLUTAMATE RECEPTOR ACTIVITY	
TRANSMEMBRANE RECEPTOR PROTEIN PHOSPHATASE ACTIVITY	
EXTRACELLULAR MATRIX PART	Related to EXTRACELLULAR MATRIX
GOLGI STACK	
PHOSPHORIC ESTER HYDROLASE ACTIVITY	
SKELETAL DEVELOPMENT	Gudbjartsson et al. (2008) Many sequence variants affecting diversity of adult human height. Nat Genet, 40, 609-615.
COLLAGEN	The most abundant proteins in ECM
ANION CATION SYMPORTER ACTIVITY	