2012 International Conference on Diabetes and Metabolism

Improving statistical powers in large scale genetic association studies

- Improving Powers in **G**enome-**W**ide **A**ssociation **S**tudies(**GWAS**)
  1. Analysis of Multiple SNPs
     - ① Regularized Regression (Elastic-Net)
     - ② Multifactor Dimensionality Reduction
     - ③ Gene-set analysis
  2. Multivariate Analysis
- T2D Consortium supported by NIDDK and preliminary analysis
- Improving Powers in **N**ext **G**eneration **S**equencing Analysis

Taesung Park

Bioinformatics and Biostatistics (BIBS) Laboratory
Department of Statistics
Seoul National University

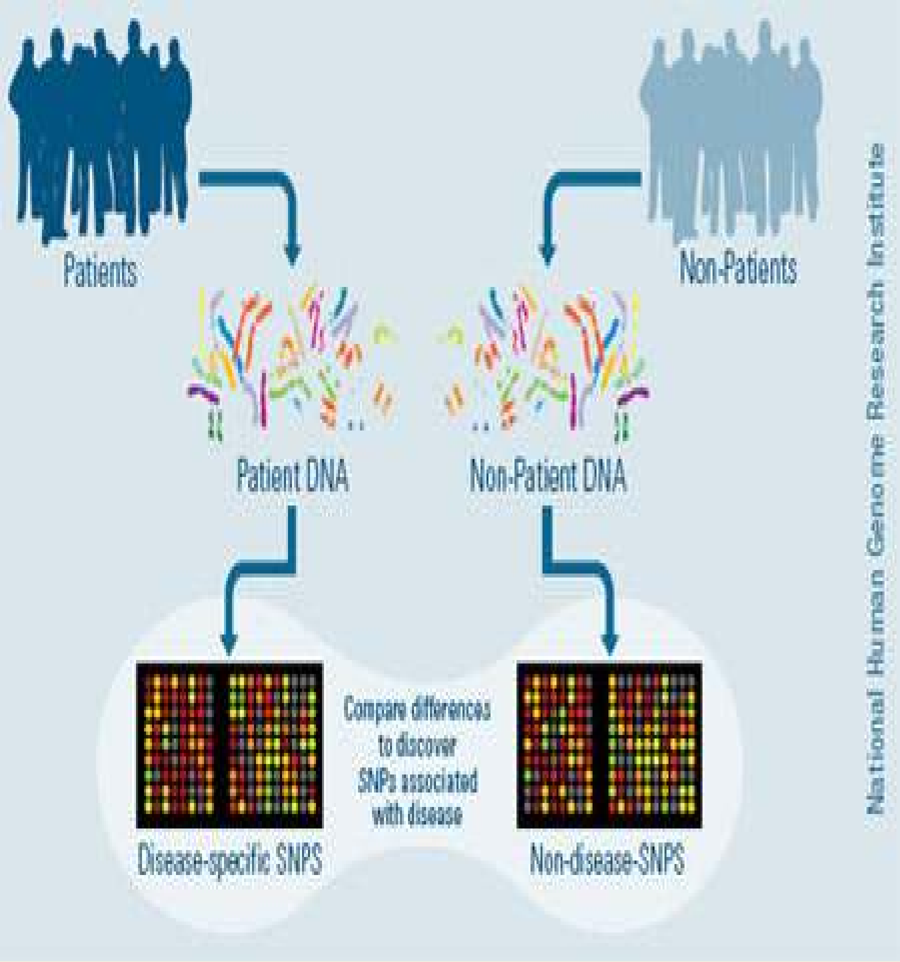
창의연구단
National Creative
Research Initiatives

Genome Wide Association Studies : GWAS

- Studies of genetic variation across the entire genome

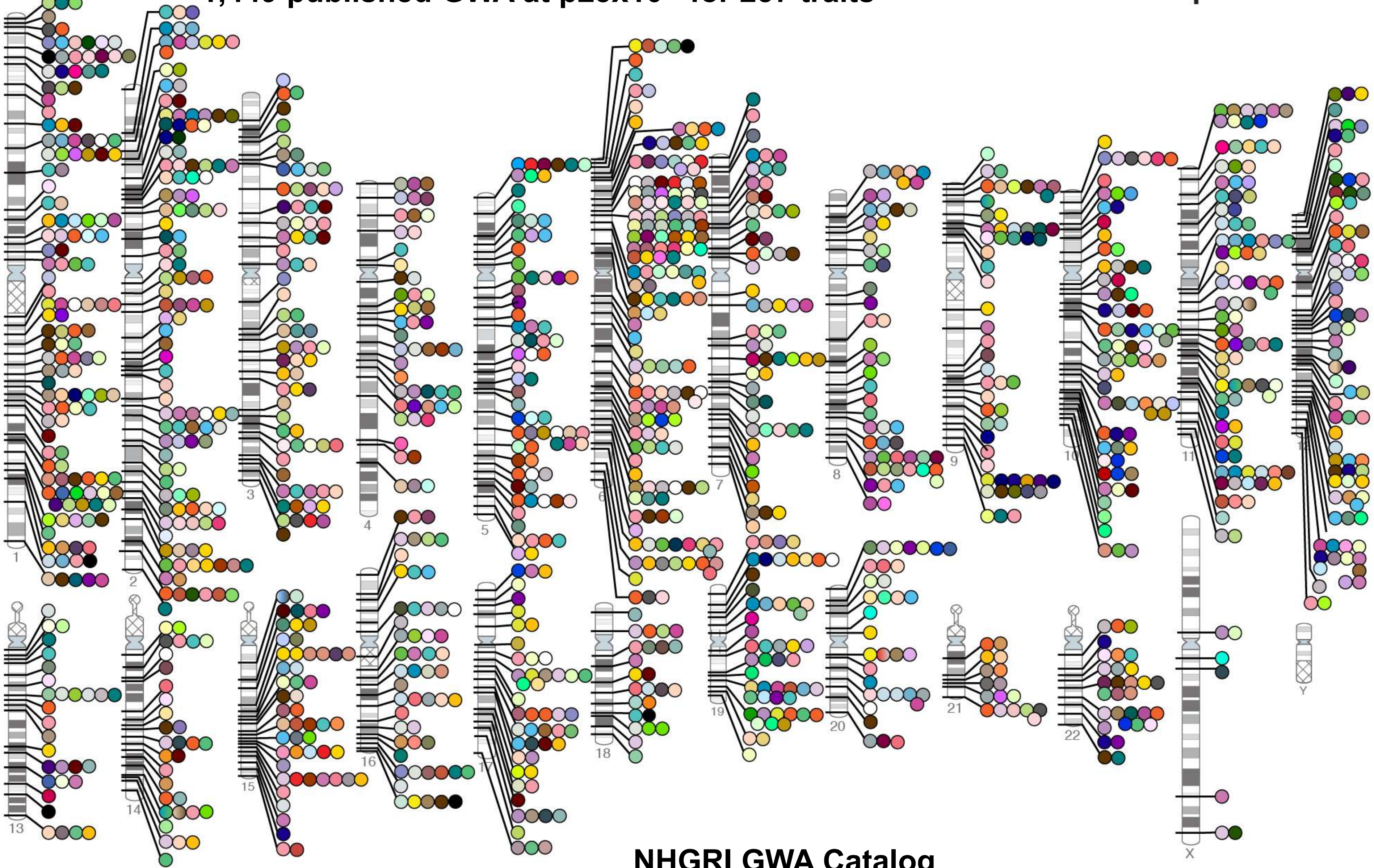
- Designed to identify associations
  between genetic markers &
  observable traits,
  or the presence/absence
  of a disease

- Rely on research tools and technologies (eg. Affy SNP chips)


Patients
Non-Patients
Patient DNA
Non-Patient DNA
Disease-specific SNPS
Compare differences to discover SNPs associated with disease
Non-disease-SNPS
National Human Genome Research Institute

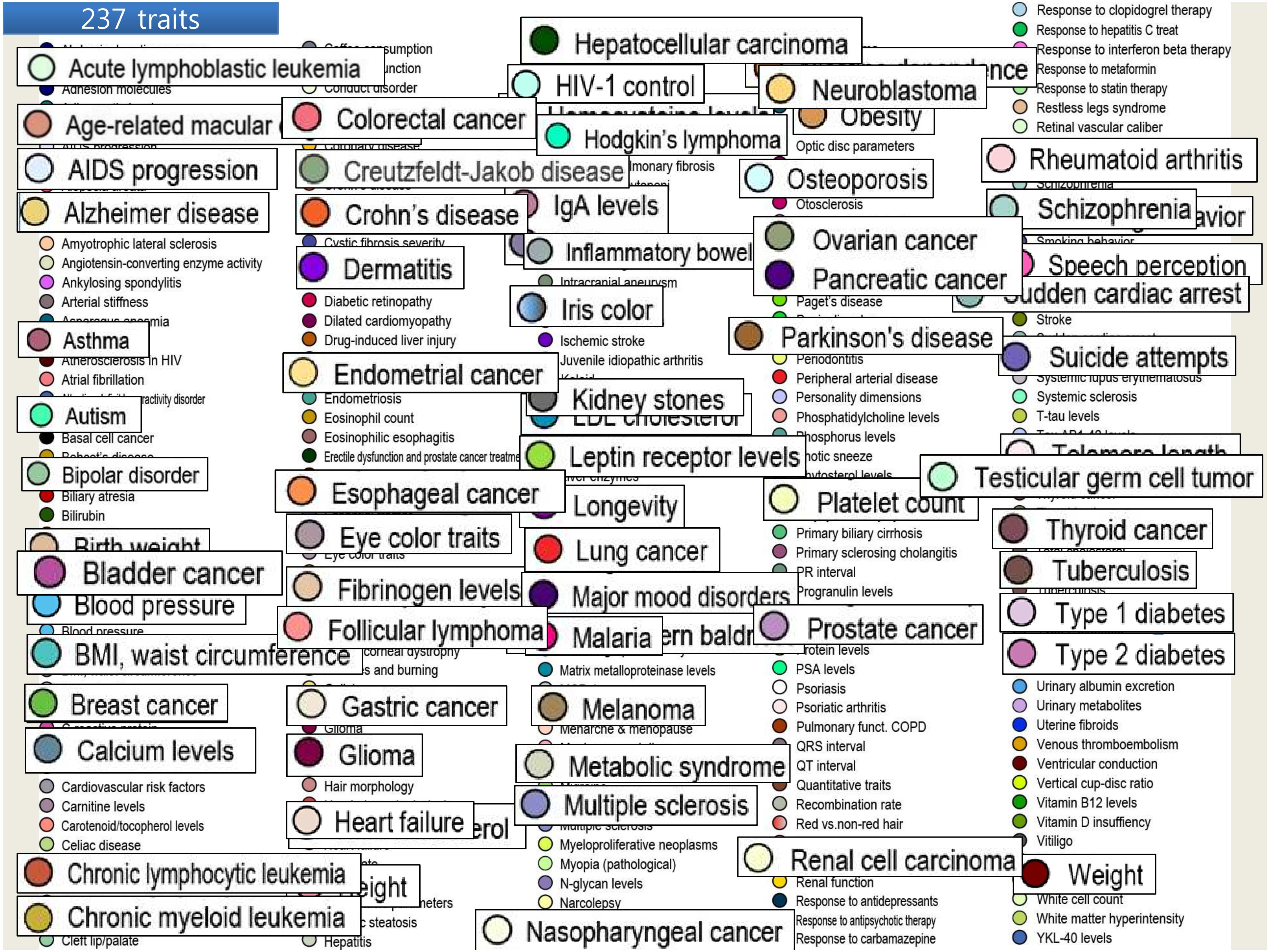
Published Genome-Wide Associations through 06/2011, 1,449 published GWA at $p \leq 5 \times 10^{-8}$ for 237 traits

2011 2nd quarter



NHGRI GWA Catalog
www.genome.gov/GWAStudies

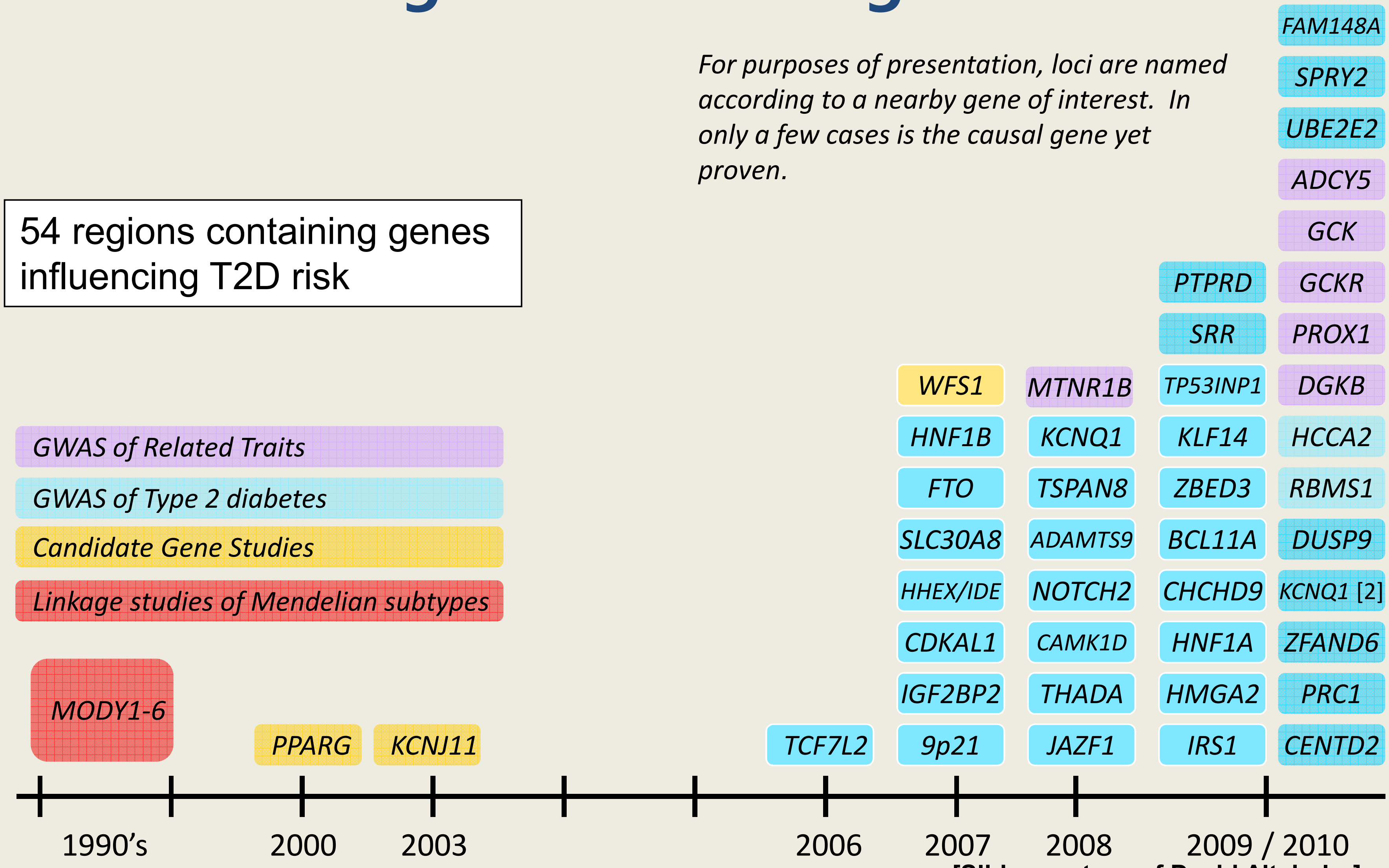
237 traits


Acute lymphoblastic leukemia
Adhesion molecules
Age-related macular
AIDS progression
AIDS progression
Alopecia areata
Alzheimer disease
Amyotrophic lateral sclerosis
Angiotensin-converting enzyme activity
Ankylosing spondylitis
Arterial stiffness
Asparagus anosmia
Asthma
Atherosclerosis in HIV
Atrial fibrillation
ractivity disorder
Autism
Basal cell cancer
Behcet's disease
Bipolar disorder
Biliary atresia
Bilirubin
Birth weight
Bladder cancer
Blood pressure
Blood pressure
BMI, waist circumference
Breast cancer
C-reactive protein
Calcium levels
Cardiovascular risk factors
Carnitine levels
Carotenoid/tocopherol levels
Celiac disease
Chronic lymphocytic leukemia
Chronic myeloid leukemia
Cleft lip/palate
Coffee consumption
unction
Conduct disorder
Colorectal cancer
Coronary disease
Creutzfeldt-Jakob disease
Crohn's disease
Crohn's disease
Cystic fibrosis severity
Dermatitis
Diabetic retinopathy
Dilated cardiomyopathy
Drug-induced liver injury
Endometrial cancer
Endometriosis
Eosinophil count
Eosinophilic esophagitis
Erectile dysfunction and prostate cancer treatme
Esophageal cancer
Eye color traits
Eye color traits
Fibrinogen levels
Follicular lymphoma
corneal dystrophy
s and burning
Gastric cancer
Glioma
Glioma
Hair morphology
Heart failure
erol
eight
meters
c steatosis
Hepatitis
Hepatocellular carcinoma
HIV-1 control
Homocysteine levels
Hodgkin's lymphoma
lmonary fibrosis
IgA levels
Inflammatory bowel
Intracranial aneurysm
Iris color
Ischemic stroke
Juvenile idiopathic arthritis
Keloid
Kidney stones
LDL cholesterol
Leptin receptor levels
iver enzymes
Longevity
Lung cancer
Major mood disorders
Malaria
ern baldn
Matrix metalloproteinase levels
Melanoma
Menarche & menopause
Metabolic syndrome
Multiple sclerosis
Multiple sclerosis
Myeloproliferative neoplasms
Myopia (pathological)
N-glycan levels
Narcolepsy
Nasopharyngeal cancer
dependence
Neuroblastoma
Obesity
Optic disc parameters
Osteoporosis
Otosclerosis
Ovarian cancer
Pancreatic cancer
Paget's disease
Parkinson's disease
Periodontitis
Peripheral arterial disease
Personality dimensions
Phosphatidylcholine levels
Phosphorus levels
hotic sneeze
hytosterol levels
Platelet count
Primary biliary cirrhosis
Primary sclerosing cholangitis
PR interval
Progranulin levels
Prostate cancer
rotein levels
PSA levels
Psoriasis
Psoriatic arthritis
Pulmonary funct. COPD
QRS interval
QT interval
Quantitative traits
Recombination rate
Red vs.non-red hair
Renal cell carcinoma
Renal function
Response to antidepressants
Response to antipsychotic therapy
Response to carbamazepine
Response to clopidogrel therapy
Response to hepatitis C treat
Response to interferon beta therapy
Response to metaformin
Response to statin therapy
Restless legs syndrome
Retinal vascular caliber
Rheumatoid arthritis
Schizophrenia
Schizophrenia
avior
Smoking behavior
Speech perception
Sudden cardiac arrest
Stroke
Suicide attempts
Systemic lupus erythematosus
Systemic sclerosis
T-tau levels
Telomere length
Testicular germ cell tumor
Thyroid cancer
Total cholesterol
Tuberculosis
Tuberculosis
Type 1 diabetes
Type 2 diabetes
Urinary albumin excretion
Urinary metabolites
Uterine fibroids
Venous thromboembolism
Ventricular conduction
Vertical cup-disc ratio
Vitamin B12 levels
Vitamin D insuffency
Vitiligo
Weight
White cell count
White matter hyperintensity
YKL-40 levels

T2D genetics through 2011

54 regions containing genes influencing T2D risk

For purposes of presentation, loci are named according to a nearby gene of interest. In only a few cases is the causal gene yet proven.



GWAS of Related Traits

GWAS of Type 2 diabetes

Candidate Gene Studies

Linkage studies of Mendelian subtypes

1990's: *MODY1-6*

2000: *PPARG*

2003: *KCNJ11*

2006: *TCF7L2*

2007: *WFS1*, *HNF1B*, *FTO*, *SLC30A8*, *HHEX/IDE*, *CDKAL1*, *IGF2BP2*, *9p21*

2008: *MTNR1B*, *KCNQ1*, *TSPAN8*, *ADAMTS9*, *NOTCH2*, *CAMK1D*, *THADA*, *JAZF1*

2009 / 2010: *PTPRD*, *SRR*, *TP53INP1*, *KLF14*, *ZBED3*, *BCL11A*, *CHCHD9*, *HNF1A*, *HMGA2*, *IRS1*

FAM148A, *SPRY2*, *UBE2E2*, *ADCY5*, *GCK*, *GCKR*, *PROX1*, *DGKB*, *HCCA2*, *RBMS1*, *DUSP9*, *KCNQ1* [2], *ZFAND6*, *PRC1*, *CENTD2*

[Slide courtesy of David Altshuler]

Korea Association Resource (KARE) Project

Objective

- To identify genetic factors of quantitative clinical traits and life-style related diseases (eg. T2DM) from Genome-Wide Association Study using population-based cohorts

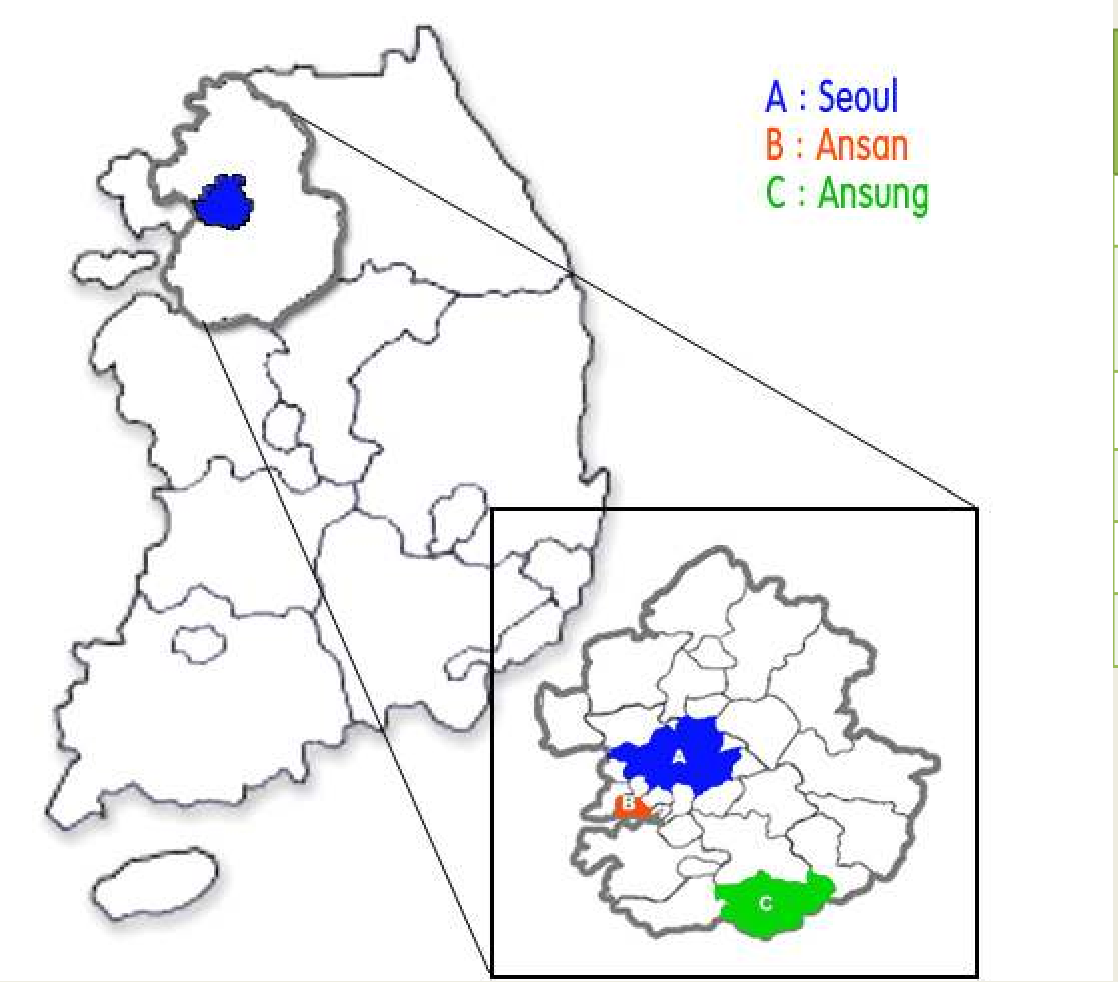
Genotyping

- Over 10,000 subjects from two community-based cohorts in Korea (Ansung & Ansan cohorts)
- Affymetrix 5.0

First high density large scale GWA Study performed in the East Asian population

Courtesy of KNIH

KARE: Characteristics

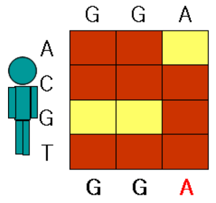


	Baseline study	
	Ansung	Ansan
Participants	5,018	5,020
Sex (women/men)	2,778/ 2,240	2,497/ 2,523
Age (mean)	55.5	49.1
40th (%)	31.2	62.8
50th (%)	29.1	23.0
60> (%)	39.6	14.3

Courtesy of KNIH

KARE: Result

SNP

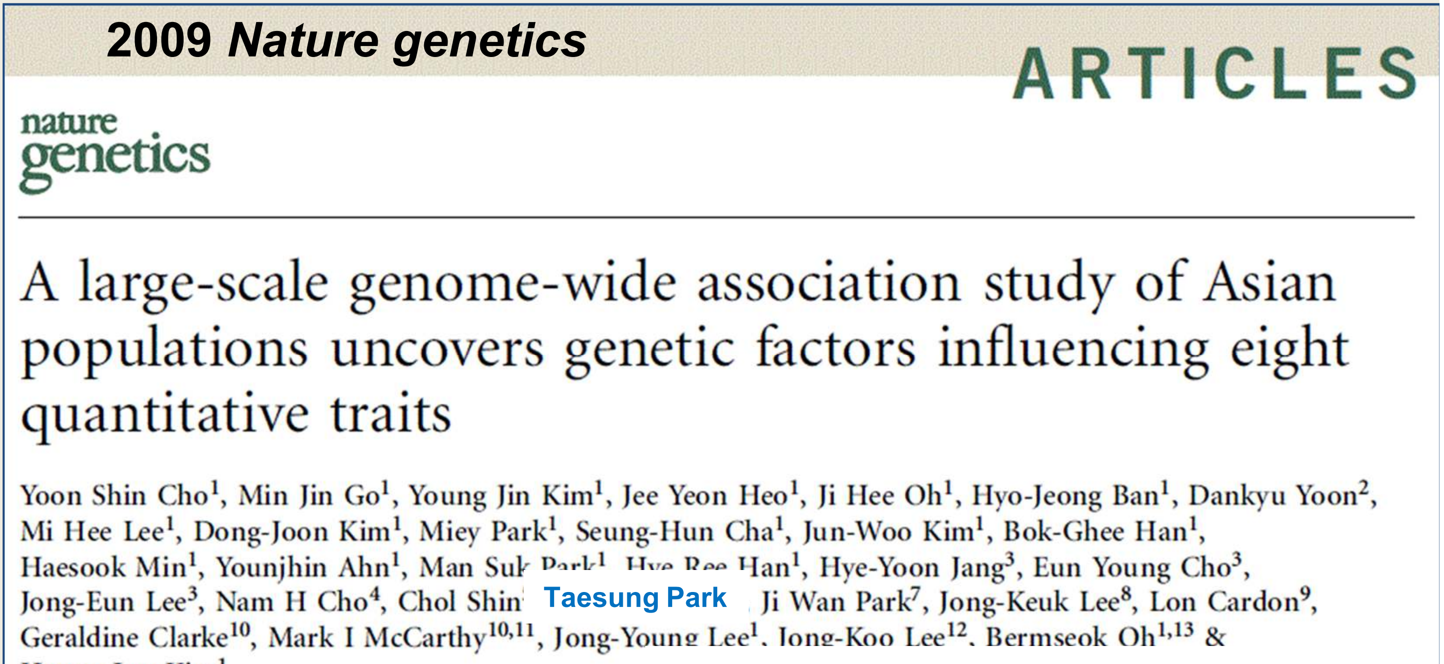


Clinical Data

Detection of 11 SNPs influencing traits in Korean population

Blood pressure, pulse rate, BMI, height, waist-hip ratio, bone mineral density

2009 *Nature genetics*

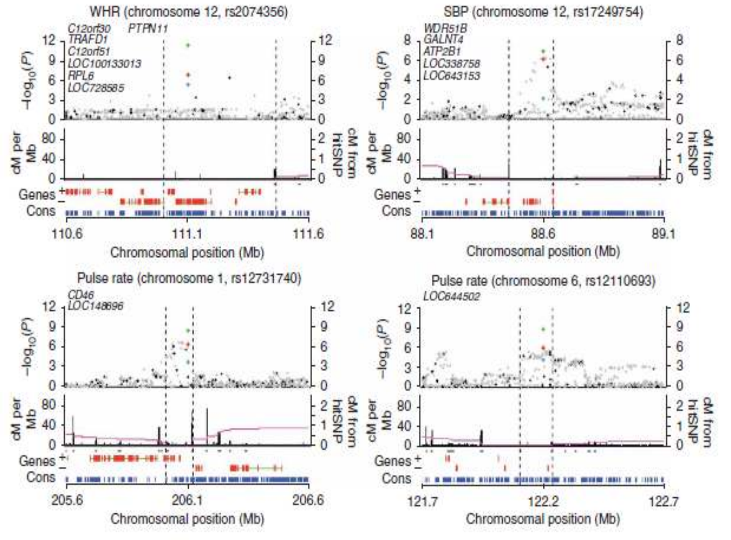


ARTICLES

nature genetics

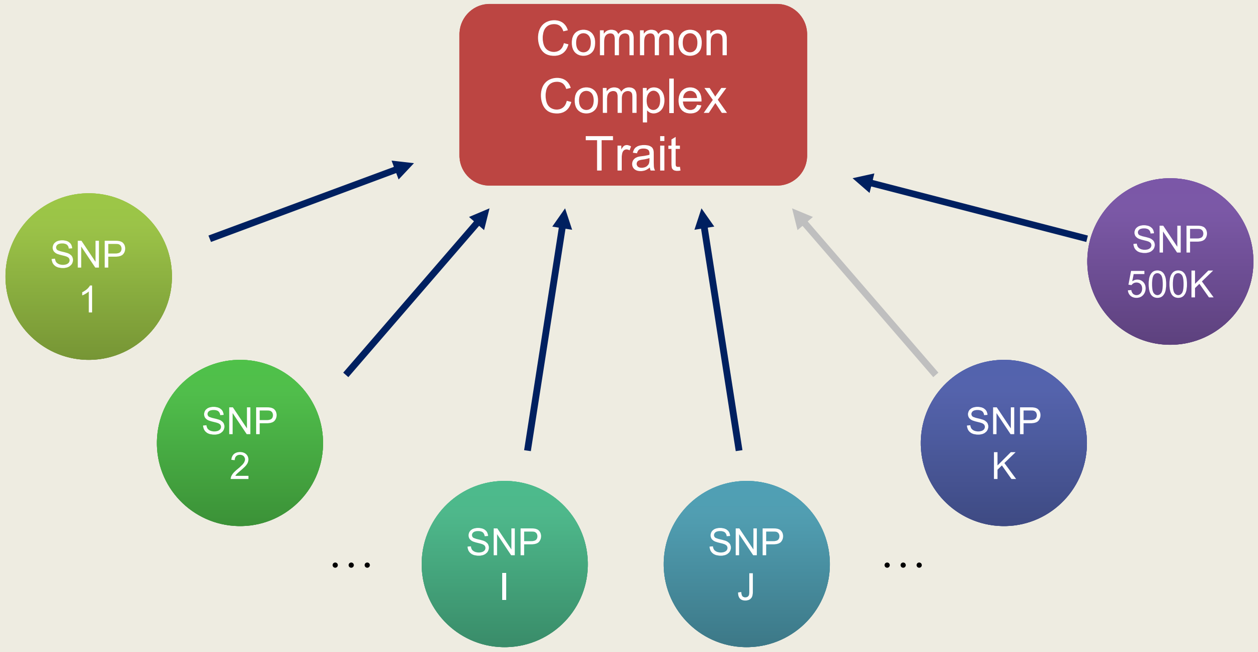
A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits

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



Current GWA Analysis

- Single SNP analysis
  - Focus on one phenotype and single SNP
  - $Trait = \beta_0 + \beta_1 SNP_i + \varepsilon$



  - Report the SNPs with high significance at $\alpha=1\times10^{-8}$

Challenges in GWAS

- Common complex traits are related with many genes
- Low power
  - Not easy to identify genetic variants with high significance at $\alpha=1\times10^{-8}$
- Not easy to get replicated results
- Further, these variants explain only small fraction of disease etiology
  - Confounding effects
  - Gene-gene and/or gene-environment interaction
- Need to develop a more powerful method for identifying genetic variants

Methods for Improving Power in GWAS

1. **Meta analysis**

2. **Analysis of multiple SNPs**

   ① Regularized Regression (Elastic-Net)

   ② Gene-Gene Interaction

   Multifactor Dimensionality Reduction

   ③ Gene Set Analysis

3. **Multivariate analysis**

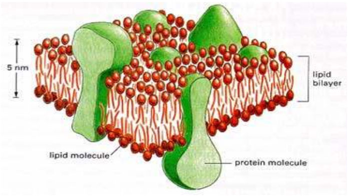
GWAS meta-analysis using KARE

European /Non-European

SNP

Clinical Data

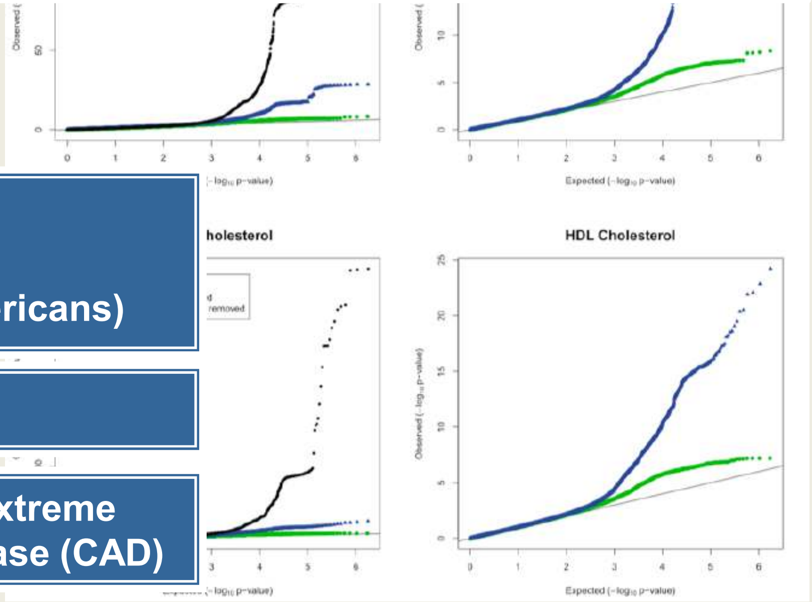
Lipid Traits



Nature, 2010

Biological, Clinical, and Population Relevance of 95 Loci Mapped for Serum Lipid Concentrations

Tanya M. Teslovich[1,118], Kiran Musunuru[2,3,4,5,6,118], Albert V. Smith[7,8], Andrew C. Edmondson[9,10], Ioannis M. Stylianou[10], Masahiro Koseki[11], James P. Pirruccello[2,5,6], Samuli Ripatti[12,13], ….. , Yoon Shin Cho[29], Min Jin Go[29], Young Jin Kim[29], Jong-Young Lee[29], **Taesung Park**[30], Kyunga J. Kim[31,32], ….. , Gonçalo R. Abecasis[1,119], **Michael Boehnke**[1,119], Sekar Kathiresan[2,3,4,5,119]



Detection of 95 loci influencing traits in 100K European population and replication study in non-European populations (East Asians, South Asians, and African Americans)

Total cholesterol (TC), LDL-C, HDL-C, TG

Identifying potential novel drug targets for treatment of extreme Lipid phenotypes and prevention of coronary artery disease (CAD)

GWAS meta-analysis using KARE

East Asian (Korea, China, Japan)

Nature Genetics, 2011

Large-scale genome-wide association studies in east Asians identify new genetic loci influencing metabolic traits

Young Jin Kim, Min Jin Go, Cheng Hu, Chang Bum Hong, Yun Kyoung Kim, , ….. , Yukinori Okada, Atsushi Takahashi, Michiaki Kubo, Toshihiro Tanaka, Naoyuki Kamatani, Koichi Matsuda, MAGIC consortium, **Taesung Park**, Bermseok Oh, Kuchan Kimm, Daehee Kang, Chol Shin, Nam H Cho, Hyung-Lae Kim, Bok-Ghee Han, Jong-Young Lee & Yoon Shin Cho

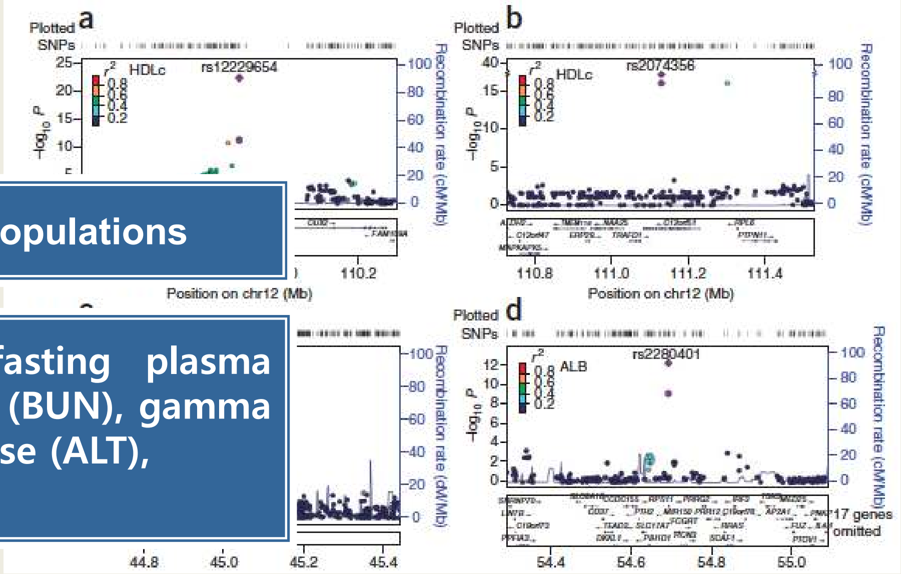
SNP

Clinical Data

Metabolic Traits

Detection of 10 loci influencing traits in east Asian populations

High density lipoprotein cholesterol (HDLc), fasting plasma glucose (FPG), albumin (ALB), blood urea nitrogen (BUN), gamma glutamyl transferase (GGT), alanine aminotransferase (ALT), aspartate aminotransferase (AST).



T2D GWAS meta-analysis using KARE

East Asian
(Korea, China, Singapore, Japan)

SNP

Clinical Data

Type 2 diabetes

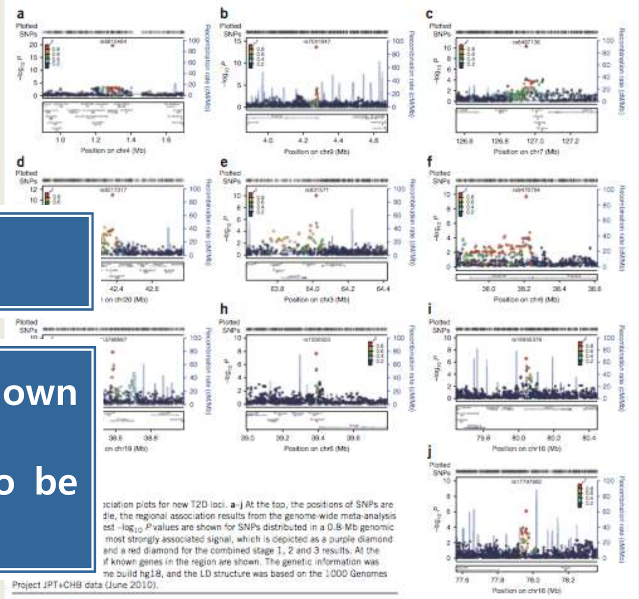
Nature Genetics, 2012

Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians

Yoon Shin Cho, Chien-Hsiun Chen, Cheng Hu, Jirong Long, Rick Twee Hee Ong, Xueling Sim, Fumihiko Takeuchi, Ying Wu, Min Jin Go, Toshimasa Yamauchi, Yi-Cheng Chang, Soo Heon Kwak, Ronald C W Ma, Ken Yamamoto, ….. Bok-Ghee Han & Mark Seielstad

Detection of 8 loci influencing T2D in east Asian populations

**Findings from this study highlight not only previously unknown biological pathways but also <u>population specific loci</u> for T2D.
The association of rs9470794 in <u>ZFAND3</u> with T2D seems to be highly specific to east Asian populations**



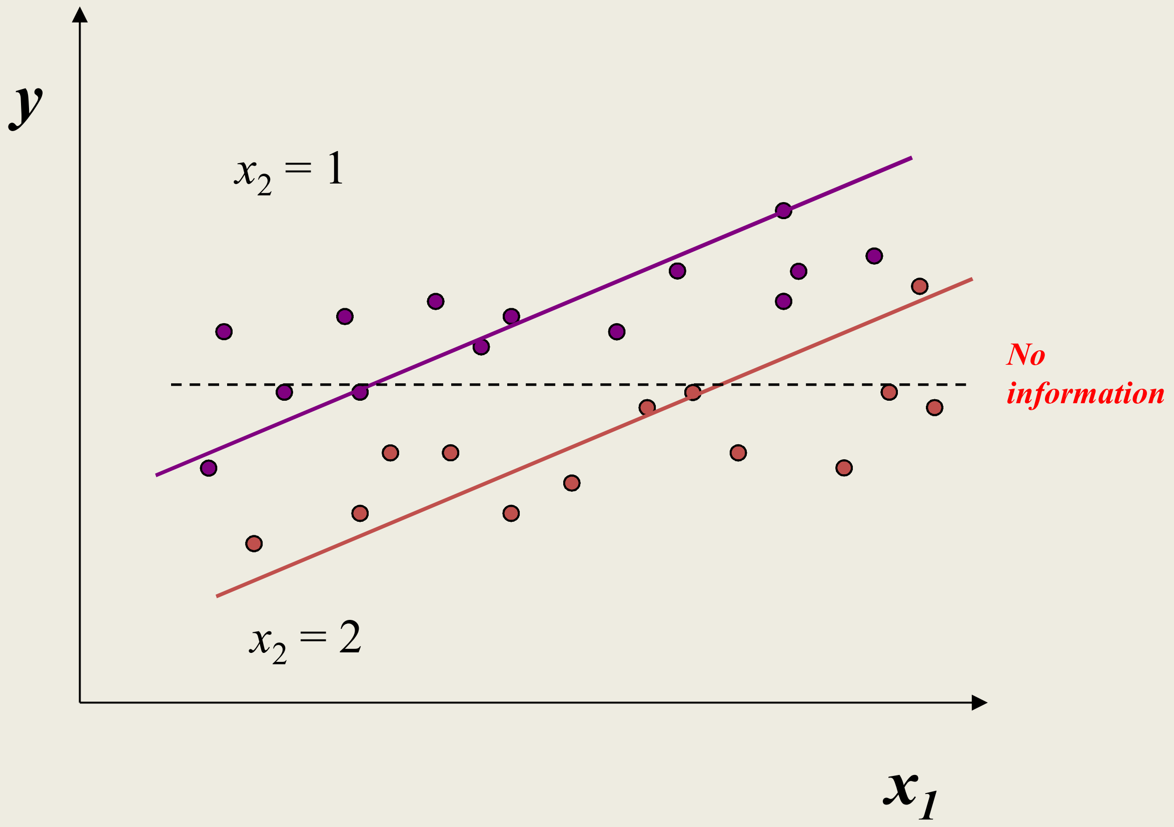
ɔciation plots for new T2D loci. **a**–**j** At the top, the positions of SNPs are
dle, the regional association results from the genome-wide meta-analysis
est $-\log_{10}$ P values are shown for SNPs distributed in a 0.8-Mb genomic
most strongly associated signal, which is depicted as a purple diamond
and a red diamond for the combined stage 1, 2 and 3 results. At the
of known genes in the region are shown. The genetic information was
ne build hg18, and the LD structure was based on the 1000 Genomes
Project JPT+CHB data (June 2010).

Improving powers in GWAS

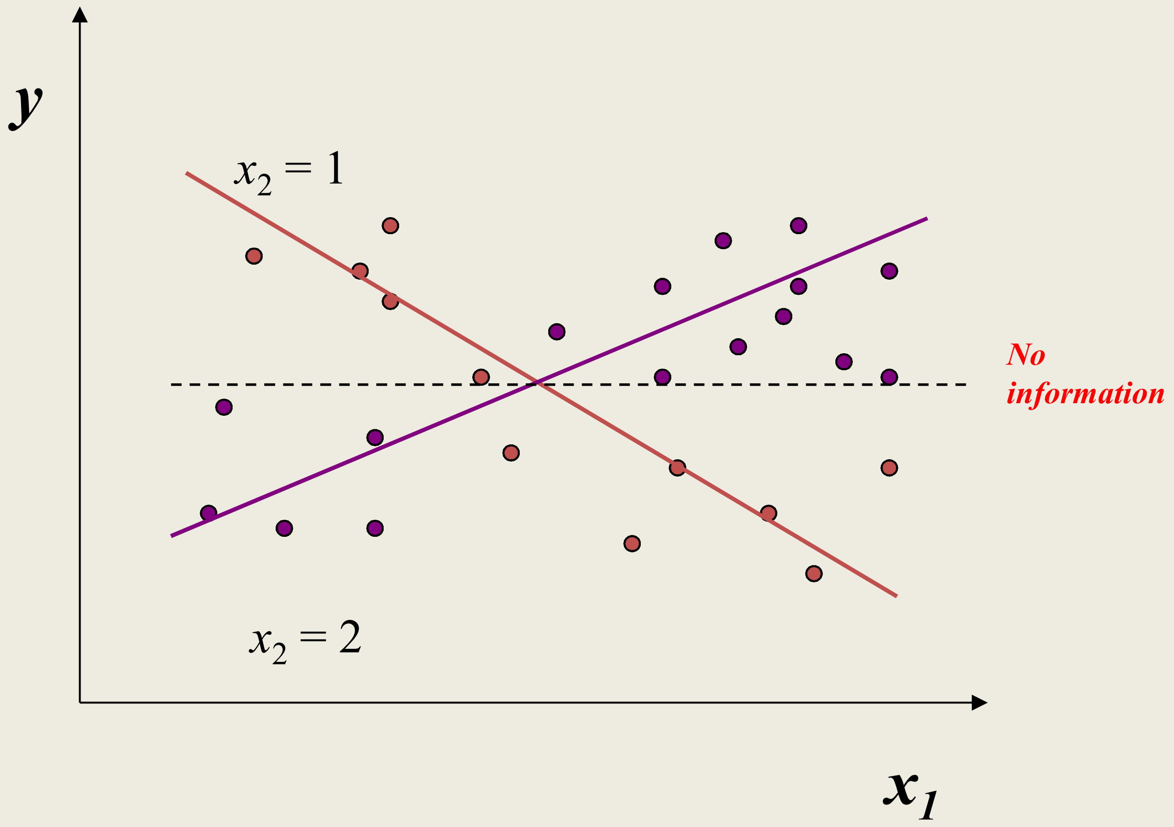
1. Meta Analysis
2. **Analysis of multiple SNPs**
   ① Regularized Regression (Elastic-Net)
   ② Gene-Gene Interaction
   Multifactor Dimensionality Reduction
   ③ Gene Set Analysis
3. Multivariate analysis

Multiple SNP Analysis

Why multiple?

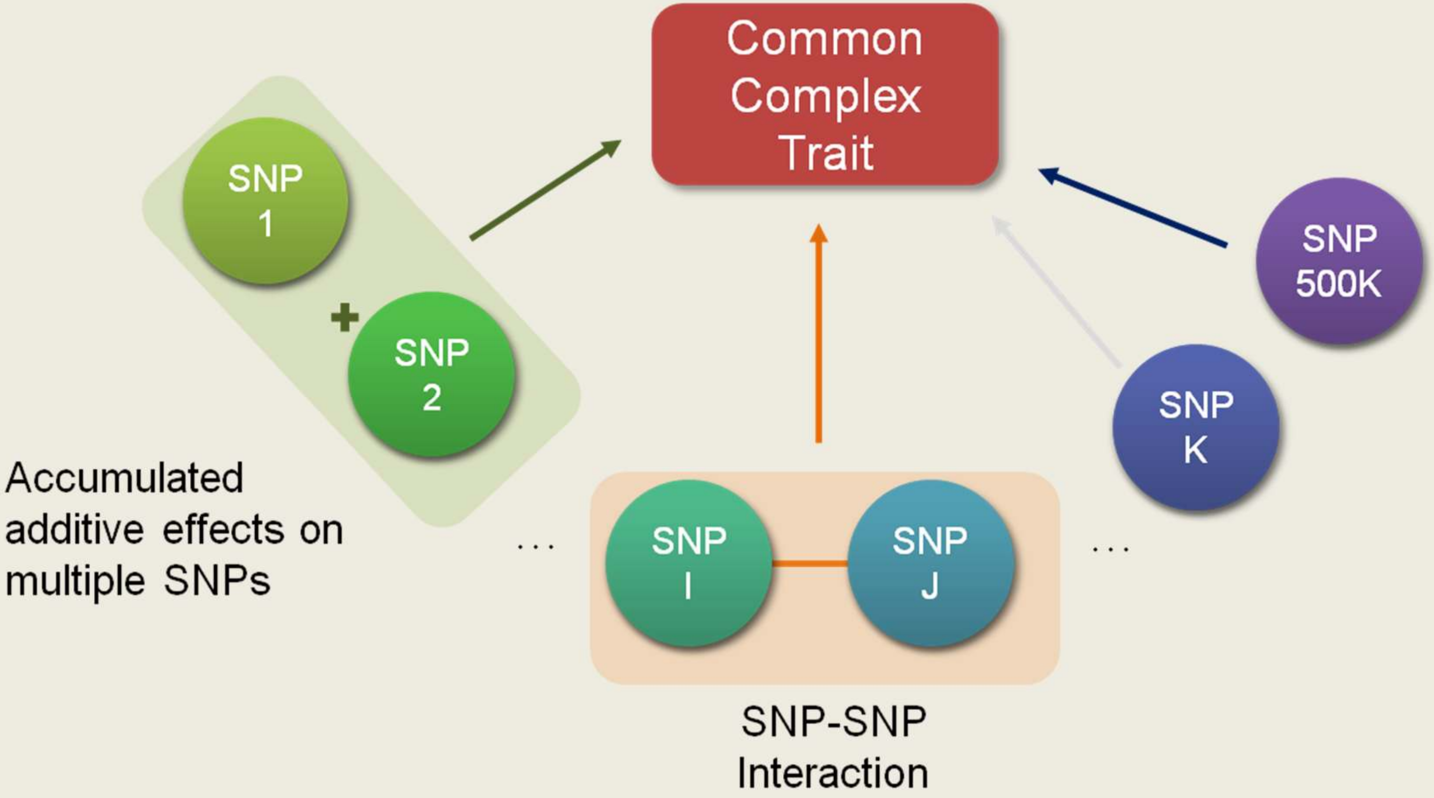


Multiple SNP Analysis
Why multiple?



Multiple SNP Analysis

- Gene-gene interaction analysis



Multiple SNP Analysis

- Current GWAS
  - Simple regression: $y_j = \beta_{0i} + \beta_i SNP_{ij} + \varepsilon_{ij} \quad (i = 1, \cdots, p, j = 1, \cdots, n)$
  - Parallel application for each of 500K SNPs

- Multiple regression
  - Model: $y_j = \beta_0 + \beta_1 SNP_{1j} + \cdots + \beta_p SNP_{pj} + \varepsilon_j \quad (j = 1, \cdots, n)$
  - High dimensionality ($n << p$): $n = 8842$, $p = 500\text{K}$
  - Correlation among input variables: LD among SNPs

Regularization

❑ Key idea: introduce ‘additional information’ to solve an ill-posed problem

- Ill-posed problems
  - Small *n*, large *p*

    n << *p* : Problem of “Curse of dimensionality”
  - Correlation among input variable

- Regularization methods
  - LASSO (L_1 penalty)
  - Ridge regression (L_2 penalty)
  - Elastic-net, composite absolute penalties

Challenges in Regularization-based variable selection

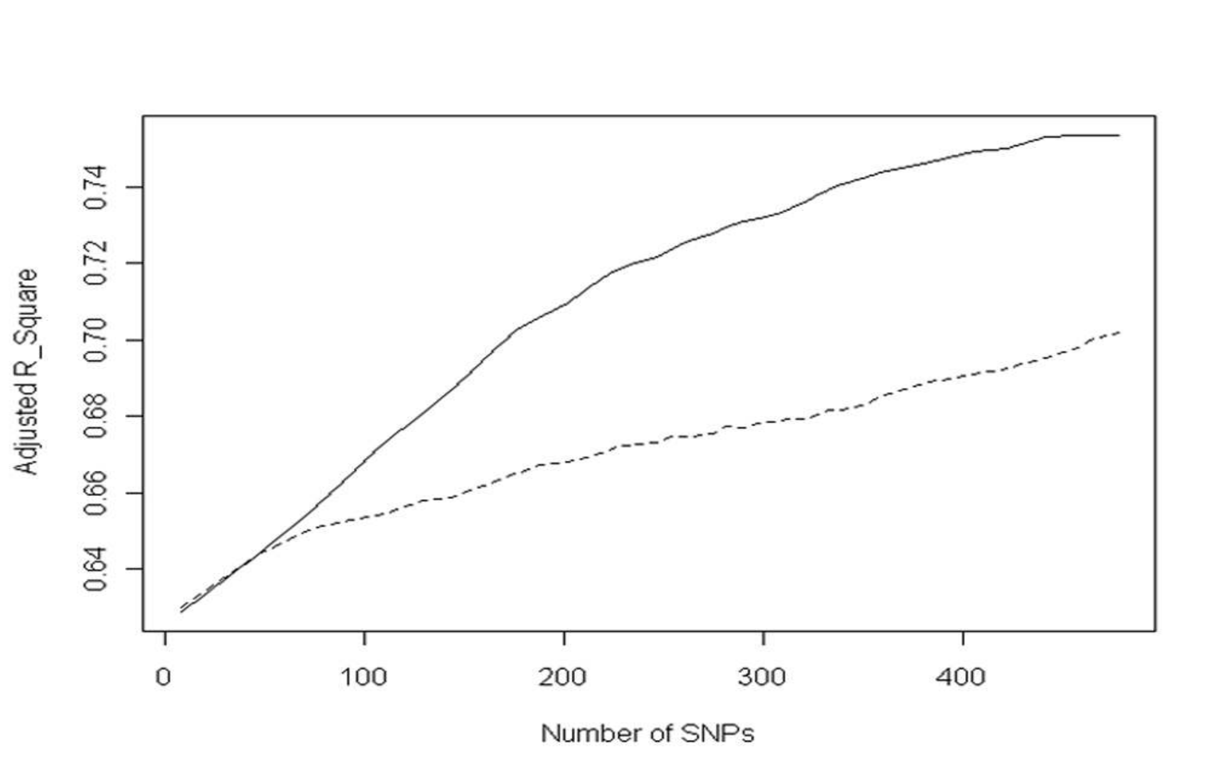
- No *p*-values for selected SNPs
  - cf. testing-based variable selection
  - Unable to provide statistical significance of selected variables
  - Hard to discuss false positives

- **Bootstrap selection stability (BSS)**
  - Generate *B* bootstrap datasets
  - Bootstrap sample is constructed by random sampling with replacement from the original dataset
  - Conduct EN variable selection with each bootstrap dataset
  - Calculate BSS for each selected SNPs

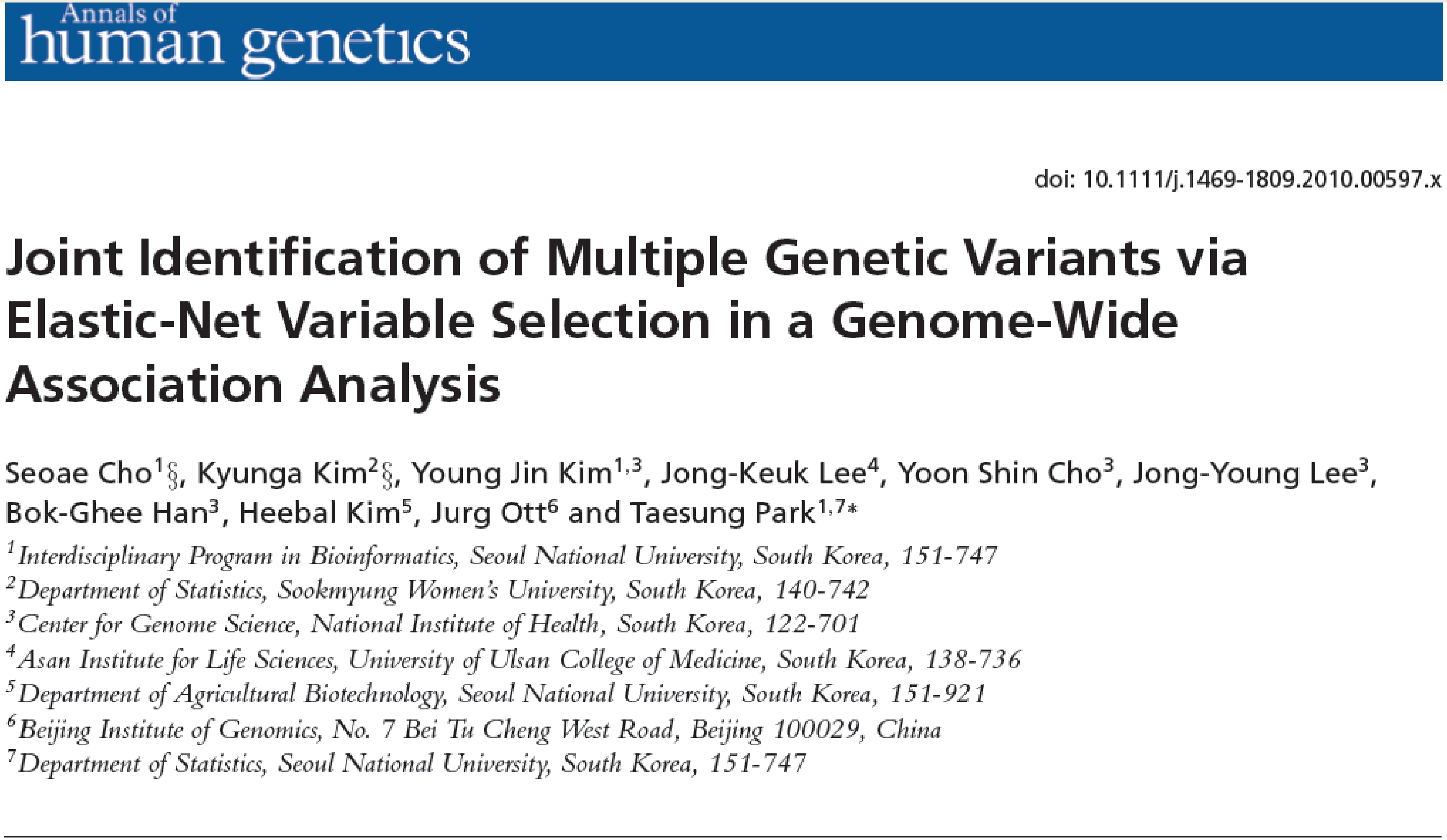
Application of EN to KARE

Explanatory Power of Identified SNPs

Proposed three-stage (solid) vs. standard (dotted)



- This power difference increased as the number of the SNPs in multiple regression models increased.



Annals of human genetics

doi: 10.1111/j.1469-1809.2010.00597.x

Joint Identification of Multiple Genetic Variants via Elastic-Net Variable Selection in a Genome-Wide Association Analysis

Seoae Cho[1]§, Kyunga Kim[2]§, Young Jin Kim[1,3], Jong-Keuk Lee[4], Yoon Shin Cho[3], Jong-Young Lee[3], Bok-Ghee Han[3], Heebal Kim[5], Jurg Ott[6] and Taesung Park[1,7]*

[1] *Interdisciplinary Program in Bioinformatics, Seoul National University, South Korea, 151-747*
[2] *Department of Statistics, Sookmyung Women's University, South Korea, 140-742*
[3] *Center for Genome Science, National Institute of Health, South Korea, 122-701*
[4] *Asan Institute for Life Sciences, University of Ulsan College of Medicine, South Korea, 138-736*
[5] *Department of Agricultural Biotechnology, Seoul National University, South Korea, 151-921*
[6] *Beijing Institute of Genomics, No. 7 Bei Tu Cheng West Road, Beijing 100029, China*
[7] *Department of Statistics, Seoul National University, South Korea, 151-747*

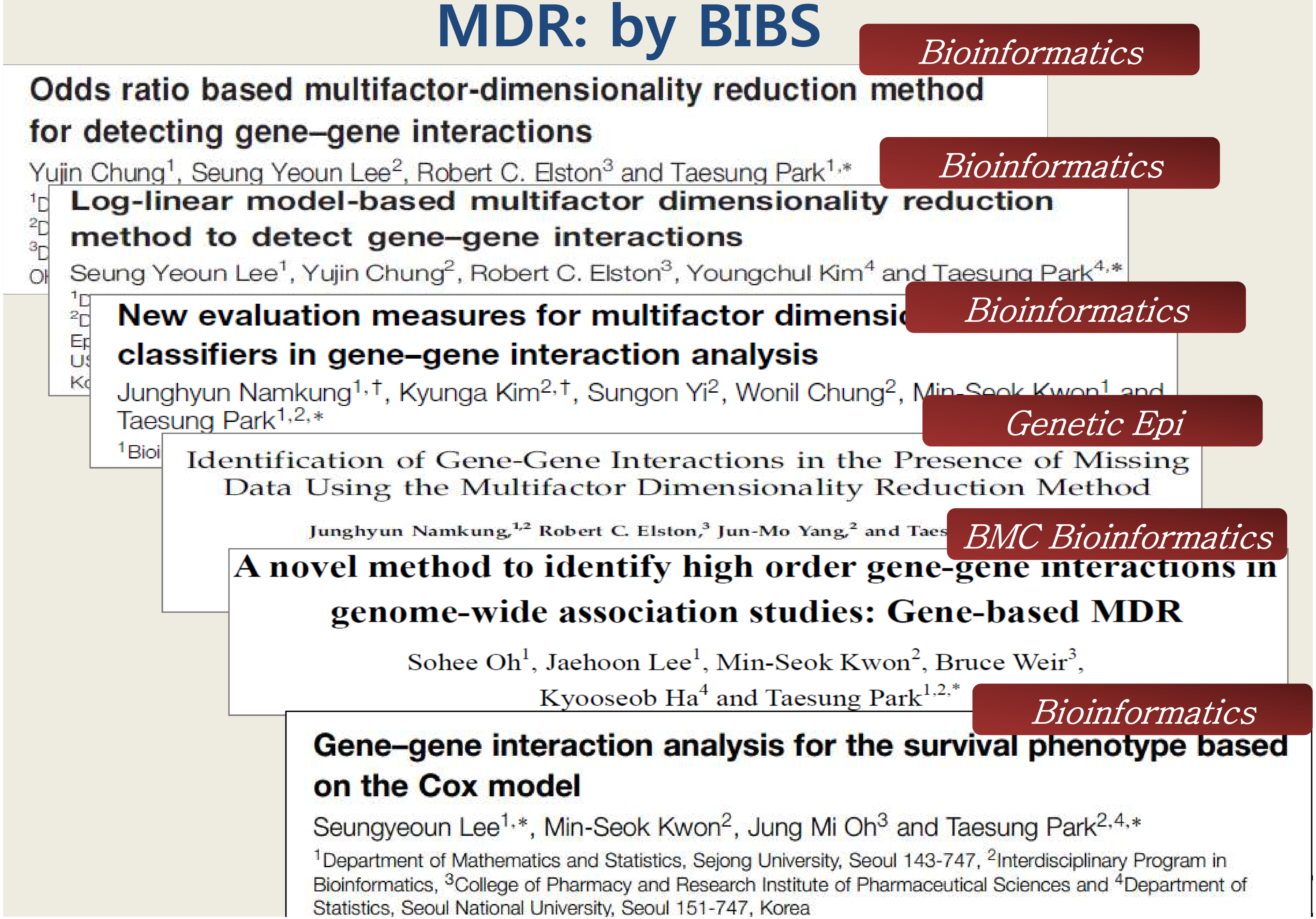
Improving powers in GWAS

1. **Analysis of multiple SNPs**
   ① Regularized Regression (Elastic-Net)
   ② Gene-Gene Interaction
   Multifactor Dimensionality Reduction
   ③ Gene Set Analysis
2. Multivariate analysis

Multifactor-Dimensionality Reduction (MDR)

- Method for detecting and characterizing interactions in common complex multifactorial disease (Ritchie *et al.*, 2001)
- Applicable even when sample size is small or dataset contains alleles in LD
- Indicate which alleles or genotypes increase susceptibility (High, Low)

MDR: by BIBS



Bioinformatics

Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions

Yujin Chung[1], Seung Yeoun Lee[2], Robert C. Elston[3] and Taesung Park[1,*]

[1]D
[2]D
[3]D
Oh

Bioinformatics

Log-linear model-based multifactor dimensionality reduction method to detect gene–gene interactions

Seung Yeoun Lee[1], Yujin Chung[2], Robert C. Elston[3], Youngchul Kim[4] and Taesung Park[4,*]

[1]D
[2]D
Ep
US
Ko

Bioinformatics

New evaluation measures for multifactor dimensi classifiers in gene–gene interaction analysis

Junghyun Namkung[1,†], Kyunga Kim[2,†], Sungon Yi[2], Wonil Chung[2], Min-Seok Kwon[1] and Taesung Park[1,2,*]

[1]Bioi

Genetic Epi

Identification of Gene-Gene Interactions in the Presence of Missing Data Using the Multifactor Dimensionality Reduction Method

Junghyun Namkung,[1,2] Robert C. Elston,[3] Jun-Mo Yang,[2] and Taes

BMC Bioinformatics

A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR

Sohee Oh[1], Jaehoon Lee[1], Min-Seok Kwon[2], Bruce Weir[3], Kyooseob Ha[4] and Taesung Park[1,2,*]

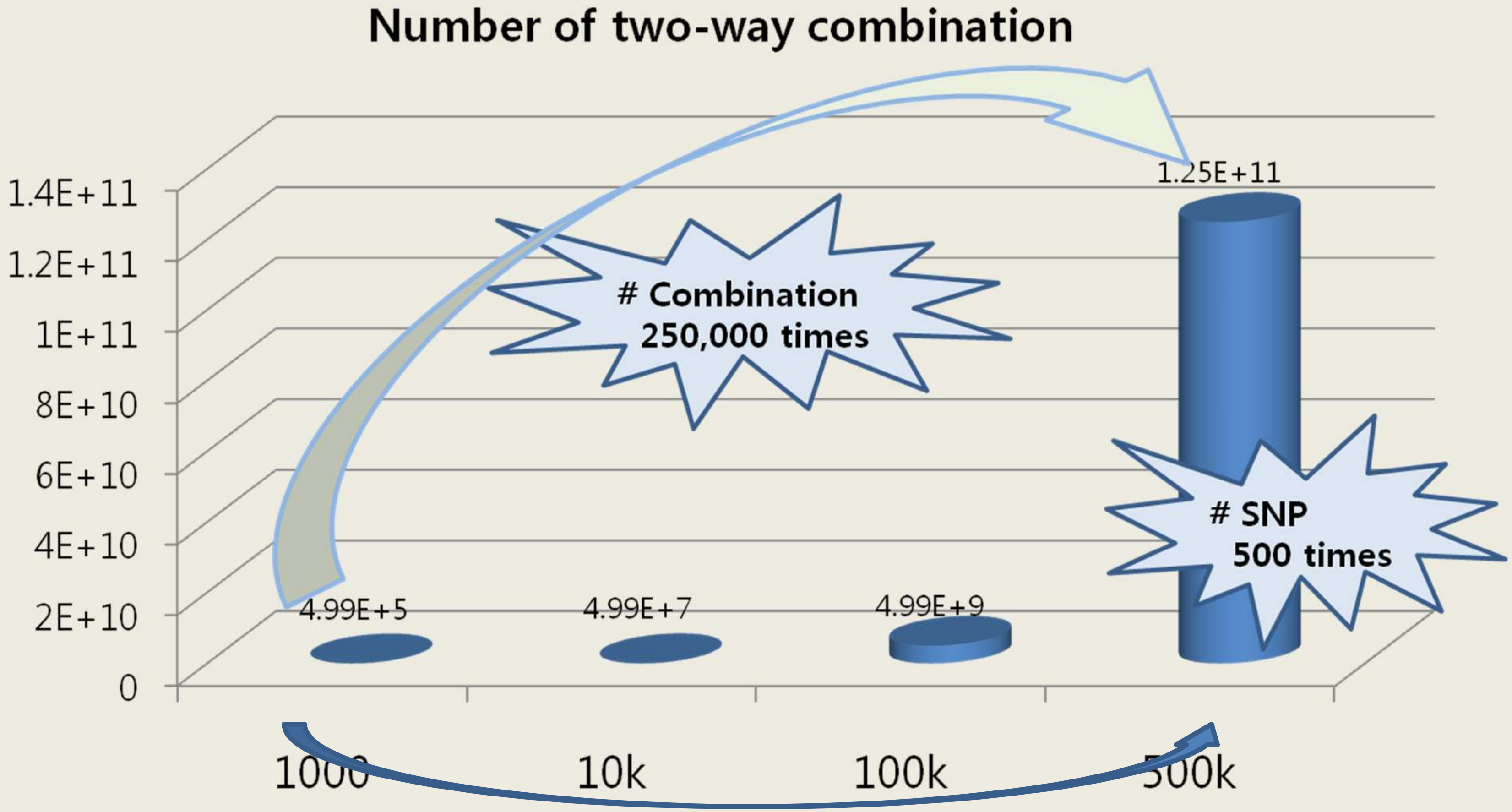
Bioinformatics

Gene–gene interaction analysis for the survival phenotype based on the Cox model

Seungyeoun Lee[1,*], Min-Seok Kwon[2], Jung Mi Oh[3] and Taesung Park[2,4,*]

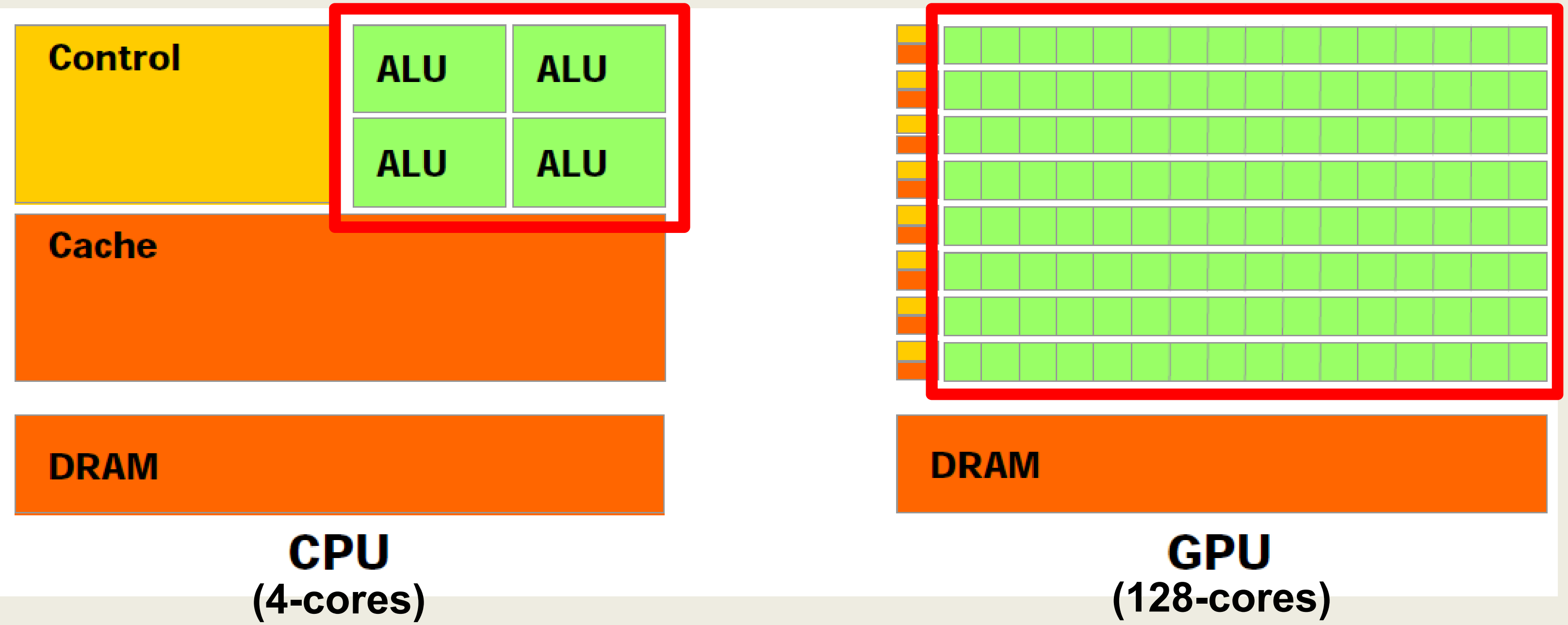
[1]Department of Mathematics and Statistics, Sejong University, Seoul 143-747, [2]Interdisciplinary Program in Bioinformatics, [3]College of Pharmacy and Research Institute of Pharmaceutical Sciences and [4]Department of Statistics, Seoul National University, Seoul 151-747, Korea

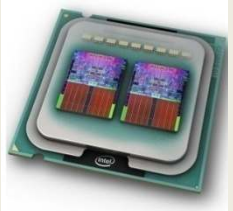
Increase in Search Space



GPU-G/MDR GPU version

Architecture of Graphic Process Unit





ALU : Arithmetic-logic unit

Performance Comparison

(CPU-based GWAS-GMDR vs. GPU-based GWAS-GMDR)

		CPU-based Computing		GPU-based Computing	
	# SNP	Xeon (1 core)	Xeon (100 cores)	3 GPU (1 node) GTX285	8 GPU (4 nodes) (= Xeon 17800 cores) Tesla M2070
SNP chip	100K	12.5 days	3 hrs	50 min	2 min
	500K	10mon	3days 3hrs	1day 20hrs	27 min
	1M	3yr 5mon	12days 11hrs	3days 9hrs	2 hrs
Reseq.	2M	13yr	1mon 19days	13days 12hrs	7 hrs
	3M	30yr	3mon 22days	1 mon	16 hrs

#sample : 1000, no. of cross-validation : 1

GTX285

Tesla M2070

Software by BIBS

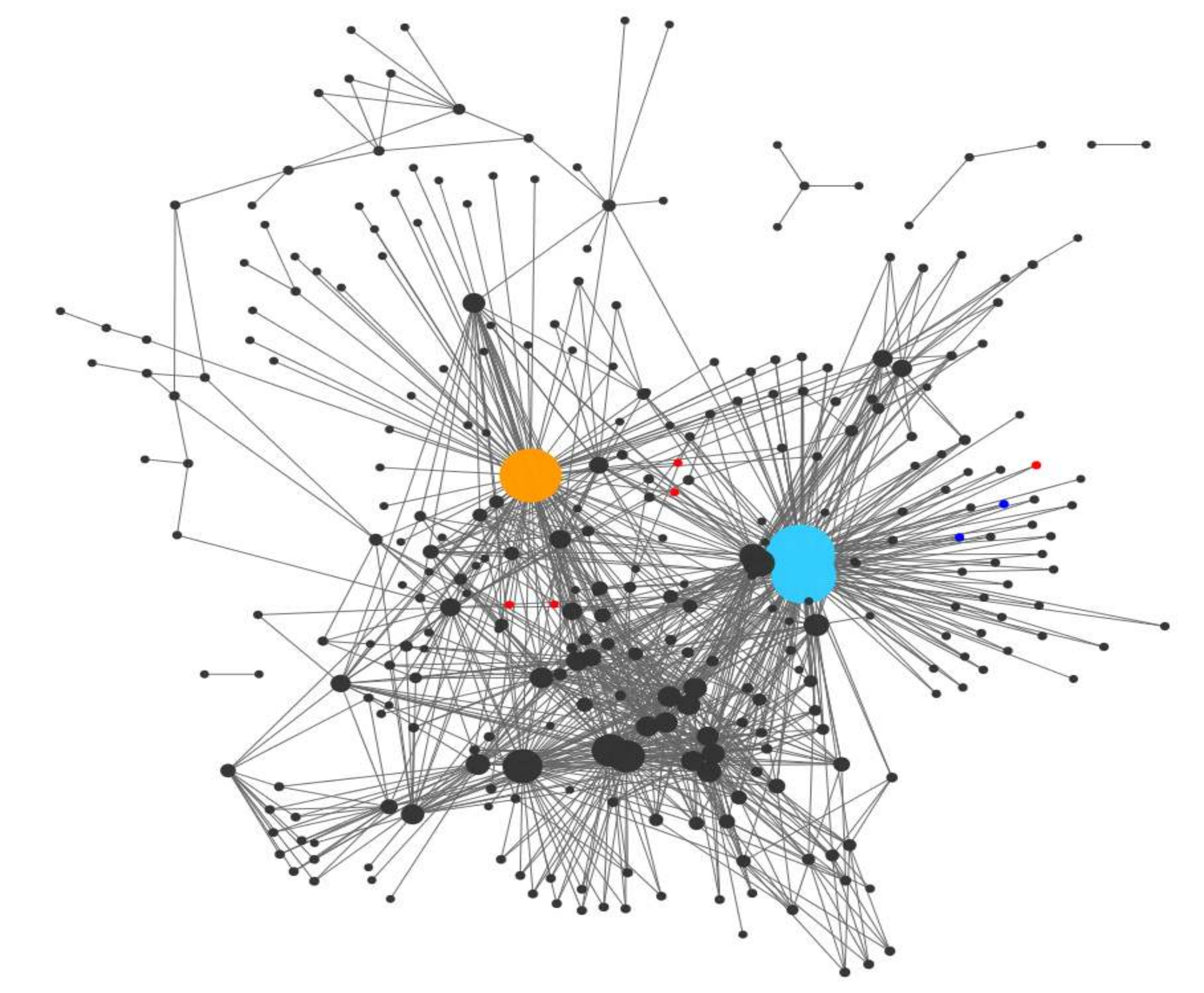
OR-MDR	Odds ratio based multifactor-dimensioinality reduction method	R package
GWAS-MDR	A program for genome-wide association analysis based on multifactor dimensionality reduction	CPU based clusters
GWAS-GMDR	A generalized GWAS-MDR that permits adjustment for covariates.	
Ordinal MDR	MDR method for ordinal phenotypes in Gene-Gene interaction analysis	
GPU-G/MDR	Ultra-high performance G/MDR program based on GPU (graphic processing unit)	GPU based system
CuGWAM	A program for visualizing gene-gene interaction in genetic association analysis	

Application of MDR to KARE

Top 20 Two-way Interactions for BMI

Rank	Best combination	WCVC	Aver. Train BA	Aver. Test BA	gene1	gene2
1	rs11590737 rs1793699	9.962234	0.577627	0.574391	PYHIN1	
2	rs1578477 rs1793699	9.925482	0.575497	0.572283		
3	rs1615480 rs1793699	9.925479	0.575497	0.572283	PYHIN1	
4	rs856127 rs1793699	9.918798	0.575109	0.573035		
5	rs7517009 rs11000212	9.898516	0.573933	0.571749	PM20D1	ASCC1
6	rs1861985 rs4921336	9.897281	0.573854	0.563459		ATP10B
7	rs4666111 rs11000212	9.896606	0.57382	0.563811	PLB1	ASCC1
8	rs2274226 rs17519968	9.888902	0.573366	0.573365	C1orf182	
9	rs1861985 rs7732722	9.885448	0.573168	0.564004		ATP10B
10	rs2274226 rs12880601	9.884424	0.573106	0.573115	C1orf182	
11	rs2597876 rs11000212	9.882513	0.573002	0.569767		ASCC1
12	rs2597875 rs11000212	9.881824	0.572963	0.568686		ASCC1
13	rs2274226 rs17519813	9.880116	0.572856	0.572859	C1orf182	
14	rs2274226 rs17441237	9.879123	0.572799	0.572803	C1orf182	
15	rs2274226 rs17441461	9.878351	0.572754	0.572757	C1orf182	
16	rs2274226 rs12434663	9.873057	0.572447	0.57245	C1orf182	
17	rs2274226 rs7147945	9.873057	0.572447	0.57245	C1orf182	
18	rs2274226 rs7146744	9.873057	0.572447	0.57245	C1orf182	
19	rs360990 rs9583489	9.871902	0.57238	0.56012		COL4A2
20	rs2274226 rs12434762	9.87129	0.572345	0.572348	C1orf182	

Two-way Interaction Network: MDR



- FTO
- FTO neighbor
- BDNF
- BDNF neighbor

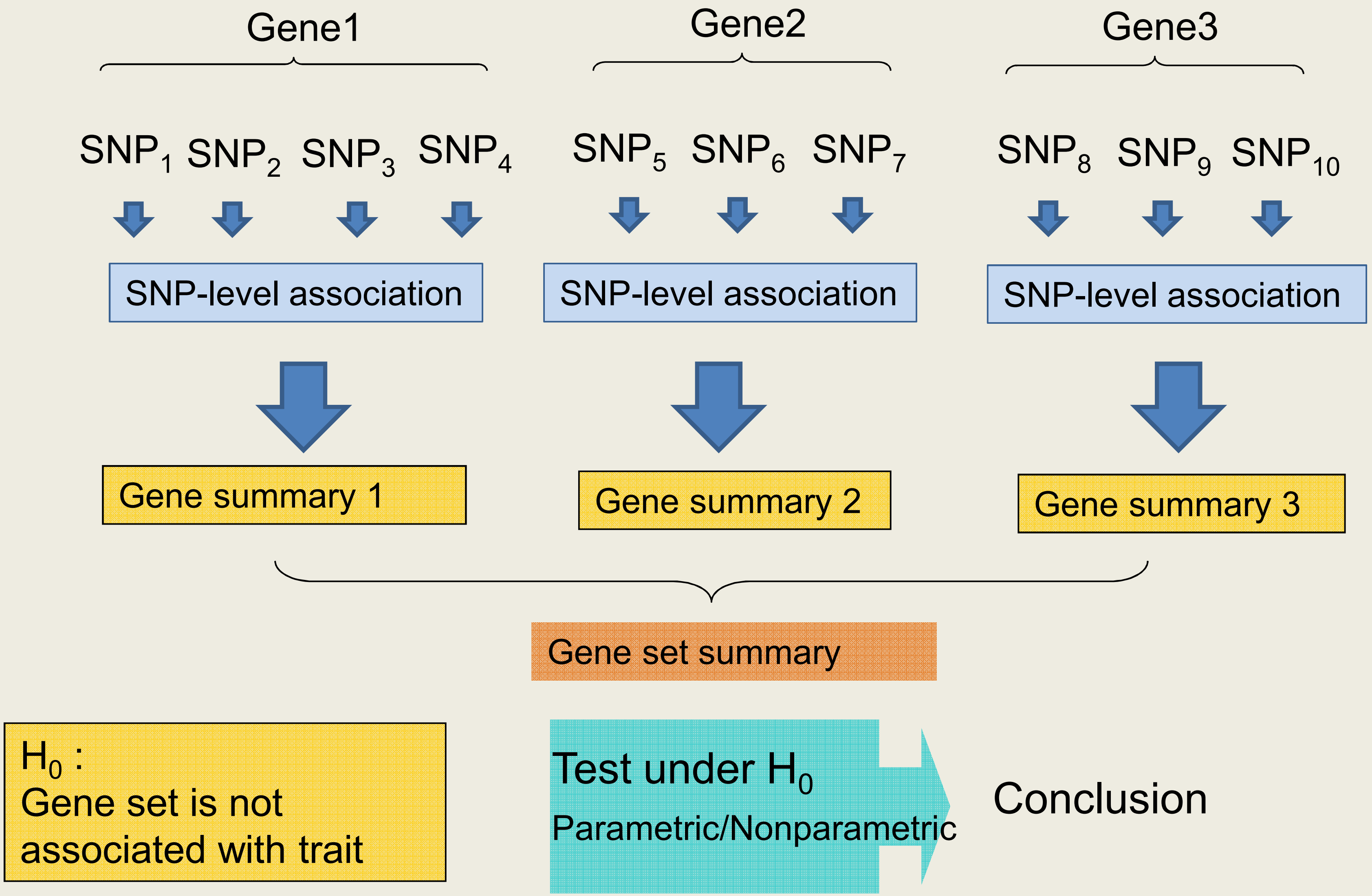
Improving powers in GWAS

1. **Analysis of multiple SNPs**
   ① Regularized Regression (Elastic-Net)
   ② Gene-Gene Interaction
   Multifactor Dimensionality Reduction
   ③ Gene Set Analysis
2. Multivariate analysis

Gene set analysis in GWAS

- Gene set
  - A pre-defined group of related genes (Biological function, Chromosomal location, regulation)
- Objective of gene set analysis (GSA)
  - Identify the gene set which is significantly associated with disease status
- Focus on gene sets rather than on individual genes or SNPs
- Benefits
  - Increase the power to detect association signals by combining weak individual signals
  - Reduce dimensionality of data
  - Provide a more expansive view of the underlying processes

Gene set analysis in GWAS



Gene1

Gene2

Gene3

SNP_1 SNP_2 SNP_3 SNP_4

SNP_5 SNP_6 SNP_7

SNP_8 SNP_9 SNP_{10}

SNP-level association

SNP-level association

SNP-level association

Gene summary 1

Gene summary 2

Gene summary 3

Gene set summary

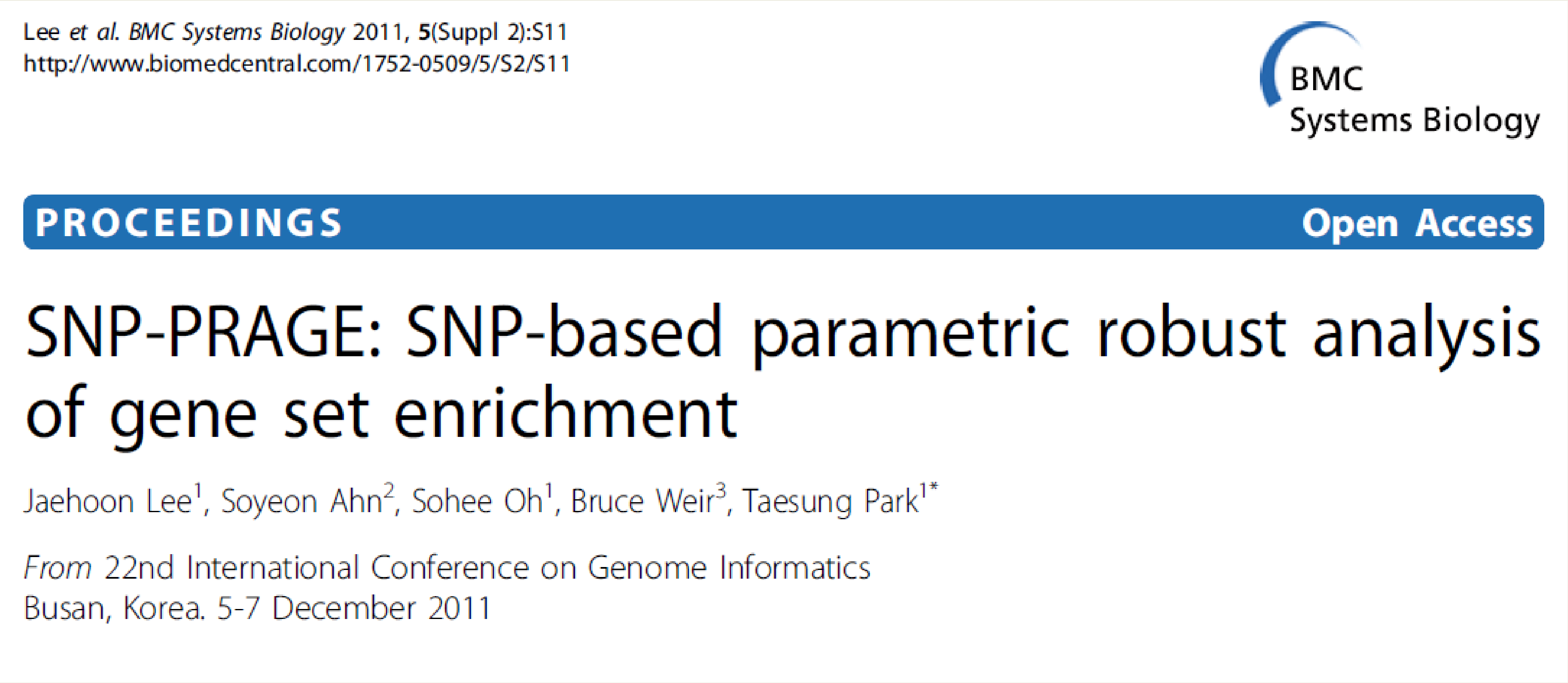
H_0 :
Gene set is not associated with trait

Test under H_0

Parametric/Nonparametric

Conclusion

Gene set analysis in GWAS



Lee *et al. BMC Systems Biology* 2011, **5**(Suppl 2):S11
http://www.biomedcentral.com/1752-0509/5/S2/S11

BMC
Systems Biology

PROCEEDINGS **Open Access**

SNP-PRAGE: SNP-based parametric robust analysis of gene set enrichment

Jaehoon Lee[1], Soyeon Ahn[2], Sohee Oh[1], Bruce Weir[3], Taesung Park[1*]

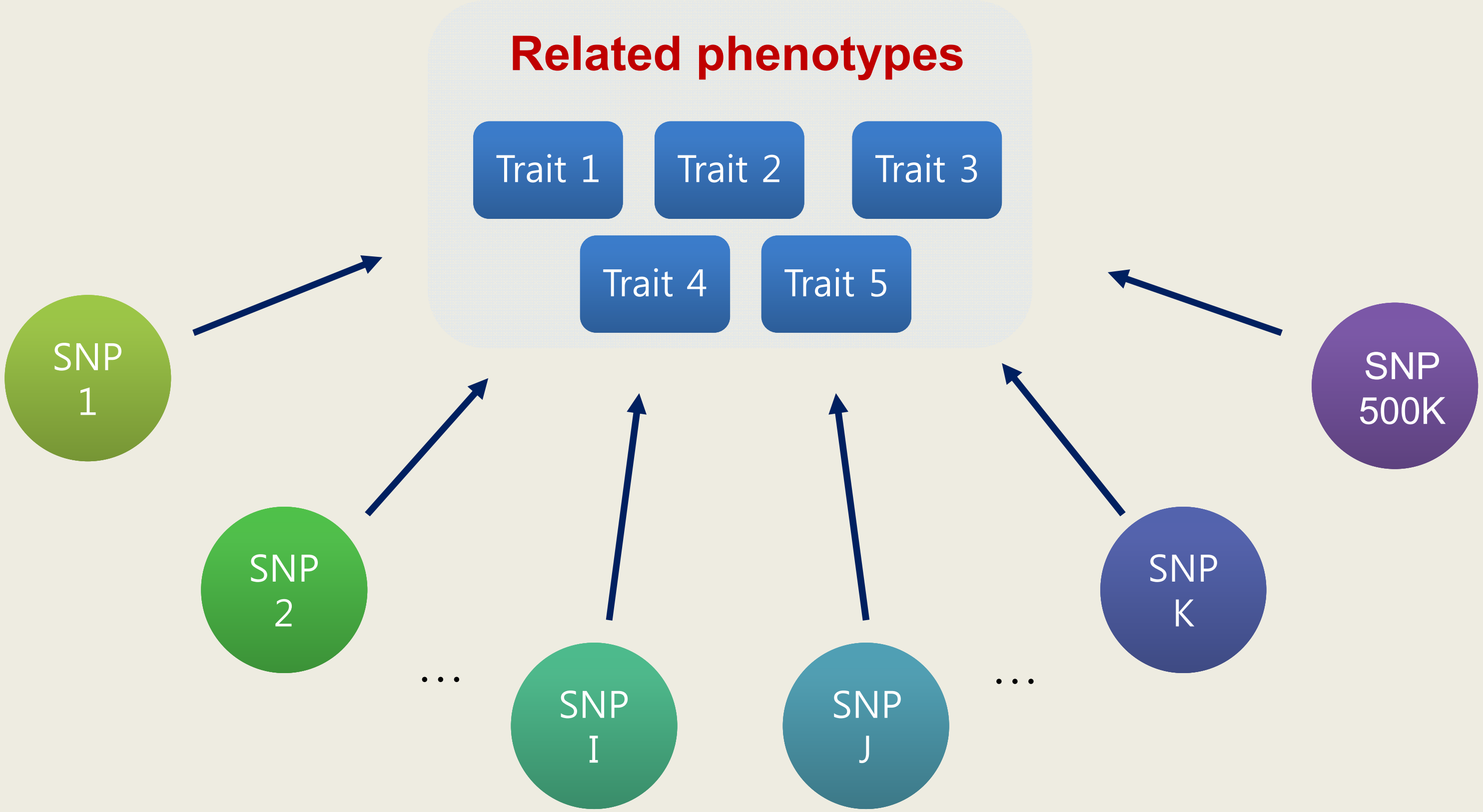
From 22nd International Conference on Genome Informatics
Busan, Korea. 5-7 December 2011

Improving powers in GWAS

1. Analysis of multiple SNPs
   ① Regularized Regression (Elastic-Net)
   ② Gene-Gene Interaction
      Multifactor Dimensionality Reduction
   ③ Gene Set Analysis
2. **Multivariate analysis**

Multivariate analysis

- Multivariate analysis
  - Focus on multiple correlated phenotypes and single SNP



Multivariate Analysis

- Examples: Related phenotypes
  - Obesity
    - BMI, Waist circumference, Weight, WHR, Body Fat
  - Hyperlipidemia
    - Total cholesterol, HDL/LDL cholesterol, Triglyceride
  - Metabolic Syndrome
    - Triglyceride, HDL cholesterol, Blood pressure, Insulin resistance

Multivariate Analysis

- Obesity related phenotypes
  - BMI, Waist circumference, Weight, and WHR
    - BMI = Weight/Height(m)^2
    - WHR = Waist / Hip circumference
  - Which genes are associated with obesity related phenotypes?

	BMI	Waist	Weight	WHR
BMI	1			
Waist	0.7607	1		
Weight	0.7308	0.6862	1	
WHR	0.3819	0.7971	0.2920	1

Univariate Analysis

- Most GWAS are conducted under this framework
- Focus on one phenotype and single SNP

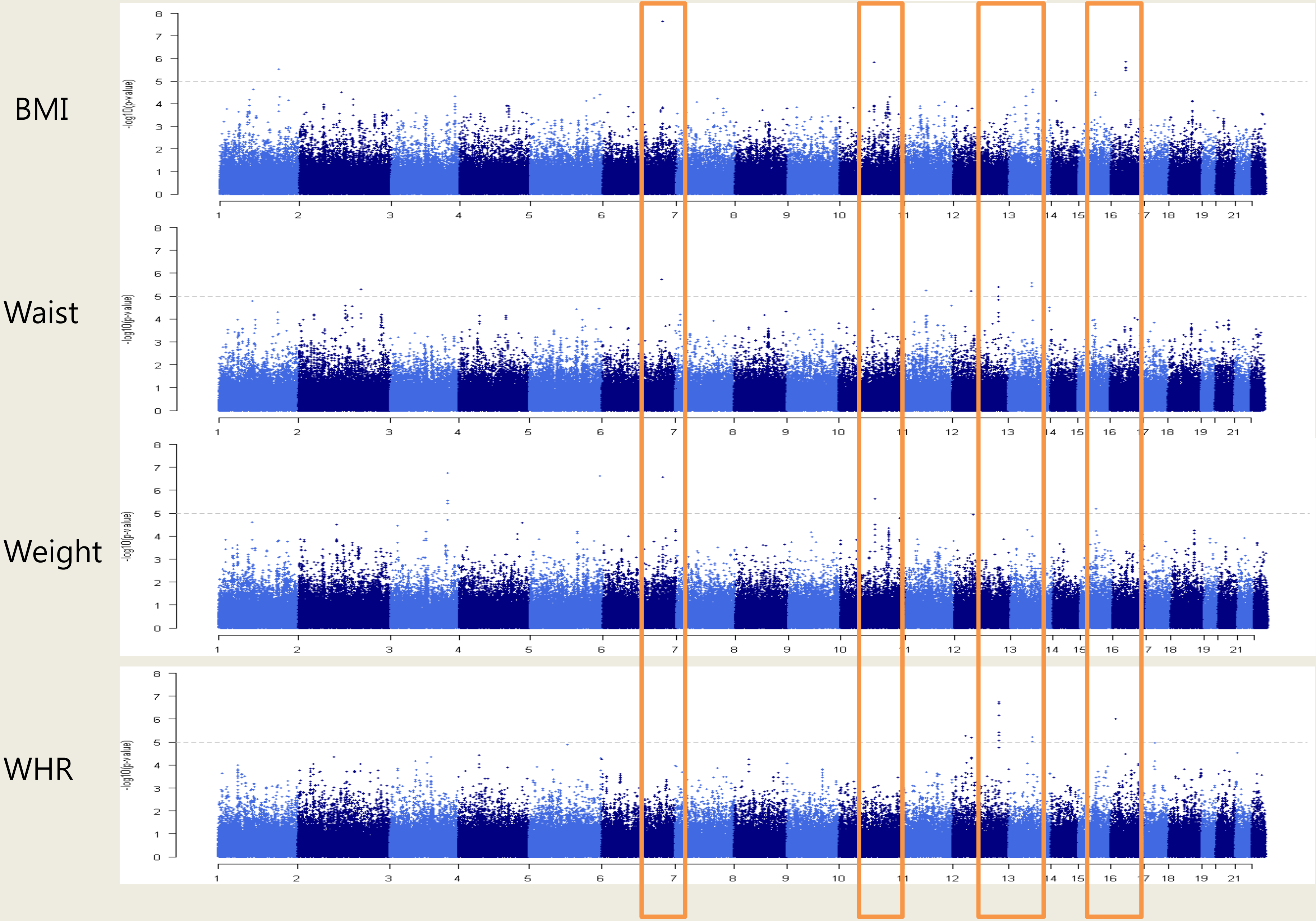
- Obesity related phenotypes
  - Separate univariate analyses

BMI:	$Y_1 = \beta_{01} + \beta_{11} Sex + \beta_{21} Age + \beta_{31} Area + \beta_{41} SNP + \varepsilon_1$
Waist:	$Y_2 = \beta_{02} + \beta_{12} Sex + \beta_{22} Age + \beta_{32} Area + \beta_{42} SNP + \varepsilon_2$
Weight:	$Y_3 = \beta_{03} + \beta_{13} Sex + \beta_{23} Age + \beta_{33} Area + \beta_{43} SNP + \varepsilon_3$
WHR:	$Y_4 = \beta_{04} + \beta_{14} Sex + \beta_{24} Age + \beta_{34} Area + \beta_{44} SNP + \varepsilon_4$

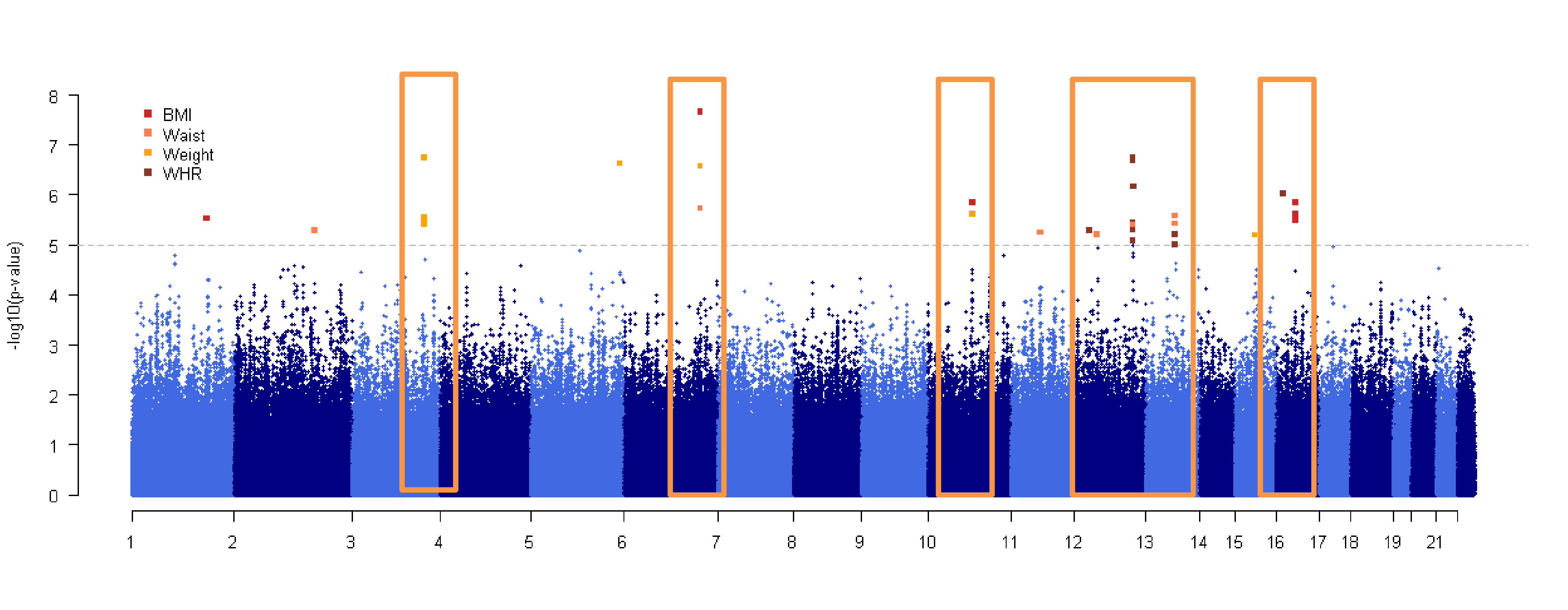
Univariate Analysis Results

- Number of significant genetic variants at a given level of α

P-value	$\leq 10^{-7}$	$10^{-7} < p \leq 10^{-6}$	$10^{-6} < p \leq 10^{-5}$	$10^{-5} < p \leq 10^{-4}$
BMI	1	0	6	23
Waist	0	0	7	39
Weight	0	3	5	32
WHR	0	4	7	25


BMI
Waist
Weight
WHR
-log10(p-value)
0
1
2
3
4
5
6
7
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
21

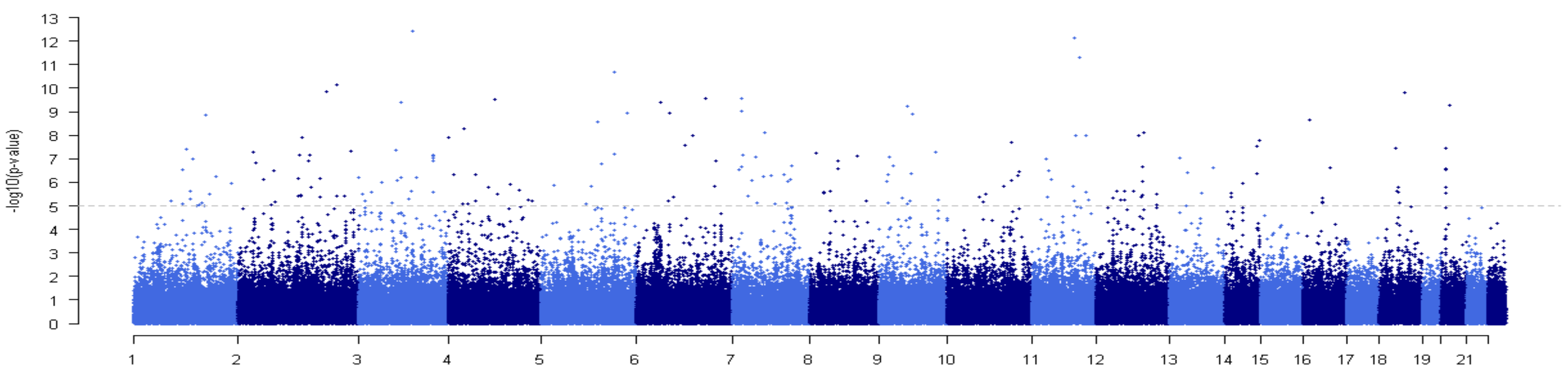
Overlay Plot



- Some SNPs have consistent significant effects on all four phenotypes

- Want to confirm by statistical testing
- Want to know whether joint analysis (multivariate analysis) of all correlated phenotypes increase power or not

Results of Multivariate Analysis



P-value	$\le 10^{-7}$	$10^{-7} < p \le 10^{-6}$	$10^{-6} < p \le 10^{-5}$	$10^{-5} < p \le 10^{-4}$
BMI	1	0	6	23
Waist	0	0	7	39
Weight	0	3	5	32
WHR	0	4	7	25
Multivariate analysis	**53**	48	89	220
	$\le 10^{-12}$	$10^{-12} < p \le 10^{-10}$	$10^{-10} < p \le 10^{-8}$	$10^{-8} < p \le 10^{-7}$
	2	3	20	28

Multivariate Analysis of KARE Data

- Newly identified obesity-related genes in KARE

CHR	SNP	P-value	Gene	Function
2	rs1377819	1.31E-08	CNTNAP5	Belongs to the neurexin family, member of which function in the vertebrate nervous system as cell adhesion molecules and receptors. This gene is related with carotid-femoral pulse wave velocity
10	rs2804219	5.24E-07	ATRNL1	A binding partner of the melanocortin-4 receptor (MC4R) gene MC4R is related to BMI and obesity and many genetic variants have been identified in GWAS
14	rs17109739	4.31E-07	NRXN3	Associated with waist circumference, BMI, and obesity

T2D Consortium

1. **Introduction**
2. **Projects**
   ① Project 1
   ② Project 2
   ③ Project 3
3. **Our preliminary analysis results**
   ① Single variant & Meta analysis
   ② Analysis of multiple variants

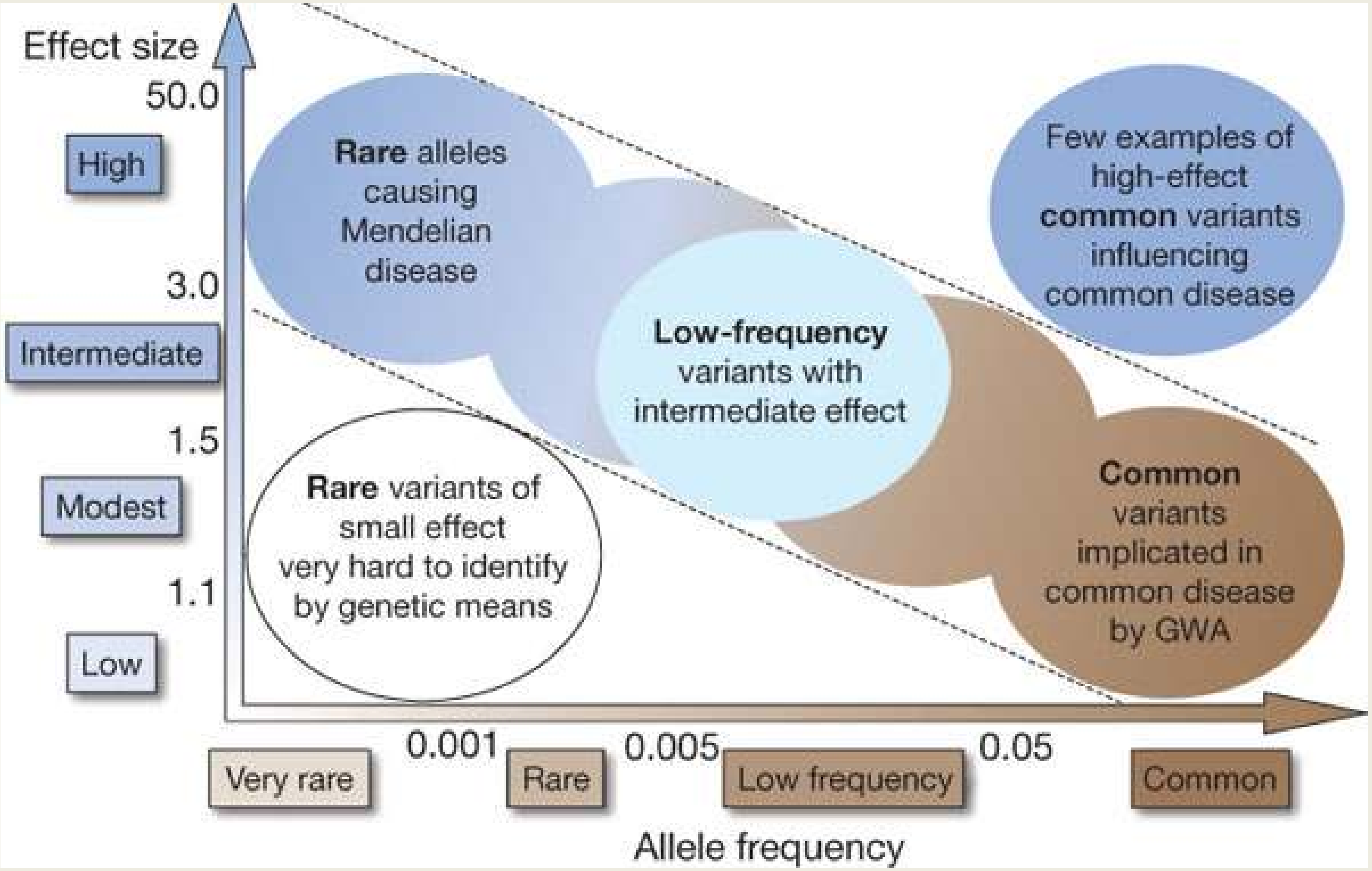
Post GWAS

- Variants identified by GWAS explain only limited proportion of genetic variability; where's the missing heritability ?
  - X chromosome
  - structural variants: indels, CNPs, CNVs

  - **G x G, G x E**
  - **less common variants with low allele frequency (<1%): => sequencing**

Post GWAS

Feasibility of identifying genetic variants


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Low-frequency variants with intermediate effect
Few examples of high-effect common variants influencing common disease
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

TA Manolio *et al. Nature* **461**, 747-753 (2009) doi:10.1038/nature08494

Motivation for T2D-Consortium

- GWAS, candidate genes have identified >60

  T2D-associated common variants

- Identified variants explain ~10% of T2D **H**^{2}

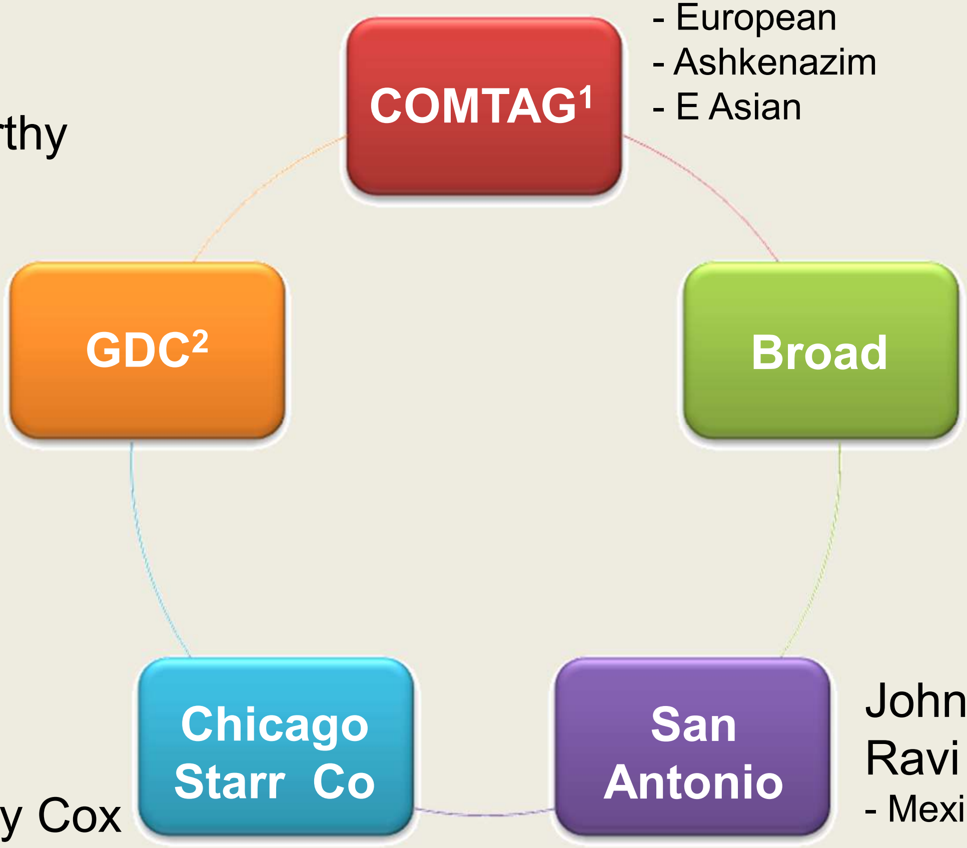
- Hypothesis :
  - ➢ less common and rare variants also contribute to T2D risk
  - ➢ may do so differentially across ancestry groups

- Large-scale sequencing studies now allow us to address this hypothesis efficiently

Introduction(cont.)

- Funding from **NIDDK** (and **NHGRI**)
- ~5 years of support: 9/20/2009 ~ 7/31/2014
- Funding :
  - $ 400-500K annual direct costs per group
  - $ 2M central funds annually

T2D Consortium



Mike Boehnke
- European
- Ashkenazim
- E Asian

David Altshuler
Jose Florez
Jim Wilson
- African American

John Biangero
Ravi Duggirala
- Mexican American

Nancy Cox
Craig Hanis
- Mexican American

Mark McCarthy
- European
- E Asian
- S Asian
- Arab
- African

- COMTAG[1] : Consortium for multiethnic type 2 diabetes associated genes
- GDC[2] : Global diabetes consortium

T2D-Consortium

- Project 1: Deep whole-exome sequencing

  (10,000 individuals from 5 ethnicities)

- Project 2: Deep whole-genome sequencing

  (600 individuals, Mexican American pedigrees)

- Project 3: Trans-ethnic fine mapping project

Investigator	Institute
Mark McCarthy	Oxford
Tim Frayling	Exeter
T Park	**SNU**
JY Lee	**KNIH**
YY Teo	Singapore
Mark Seielstad	UCSF
Mike Boehnke	Michigan University
Rob Sladek	Montreal

T2D Consortium

1. Introduction
2. **Projects**
   ① Project 1
   ② Project 2
   ③ Project 3
3. Our preliminary analysis results
   ① Single variant & Meta analysis
   ② Analysis of multiple variants

Project 1 : Introduction

- Project 1 seeks to assess whether less common variants play a role in T2D risk and to assess similarities and differences in the distribution of T2D risk variants across ancestry groups.
- Five ancestry groups : European, East Asians, South Asians, American Hispanics, and African Americans.
- Sequencing is underway at the Broad using the Agilent v2 capture reagent on Hiseq machines(65x coverage).

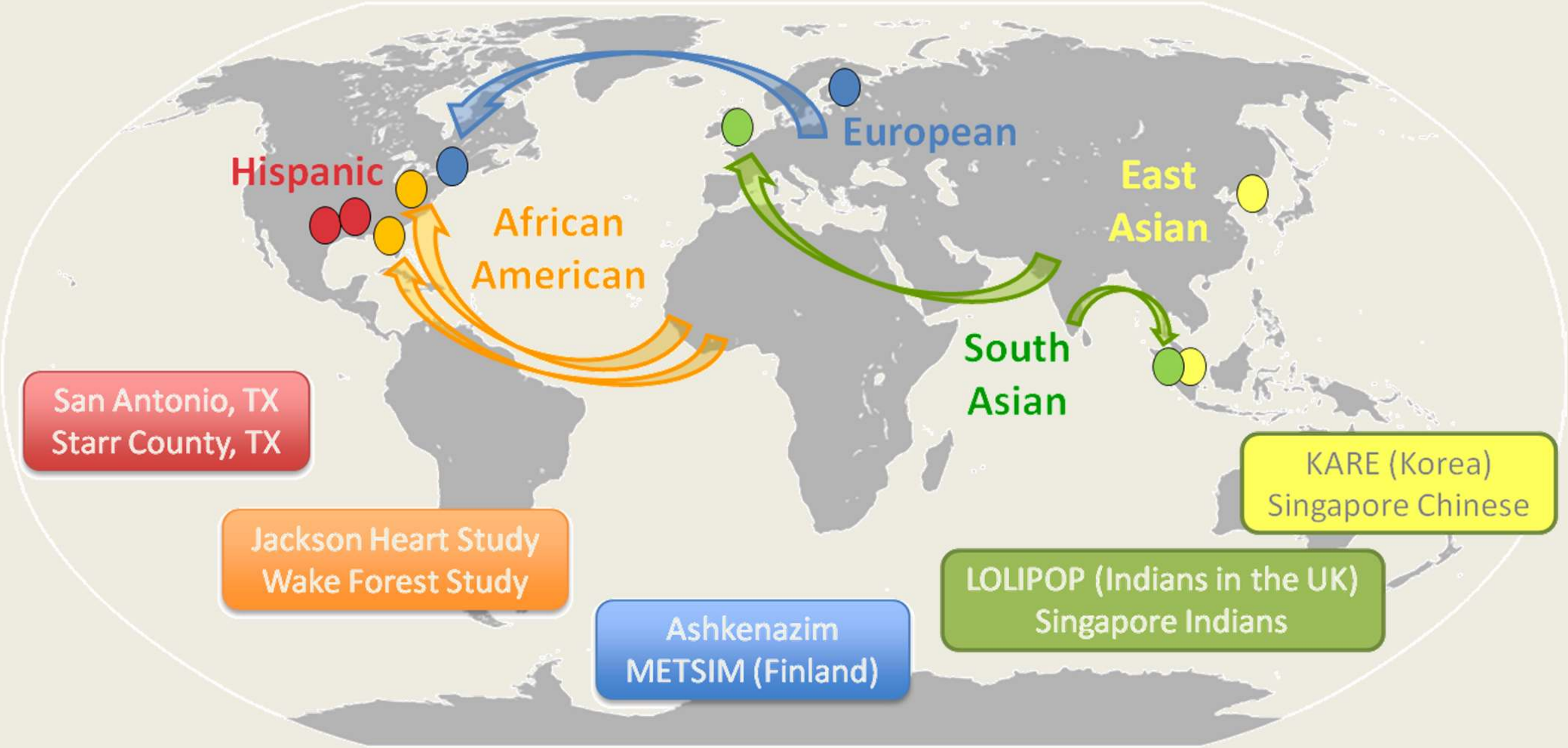
Samples selected for sequencing

- 500 cases / 500 controls from each of 10 cohorts from 5 ethnicities

Population	Study description
African American	Jackson Heart Study Wake Forest
East Asian	Korean Chinese from Singapore
European	Ashkenazi Finnish (METSIM)
Hispanic	San Antonio Starr County
South Asian	Indians living in London (LOLIPOP*) Indians from Singapore

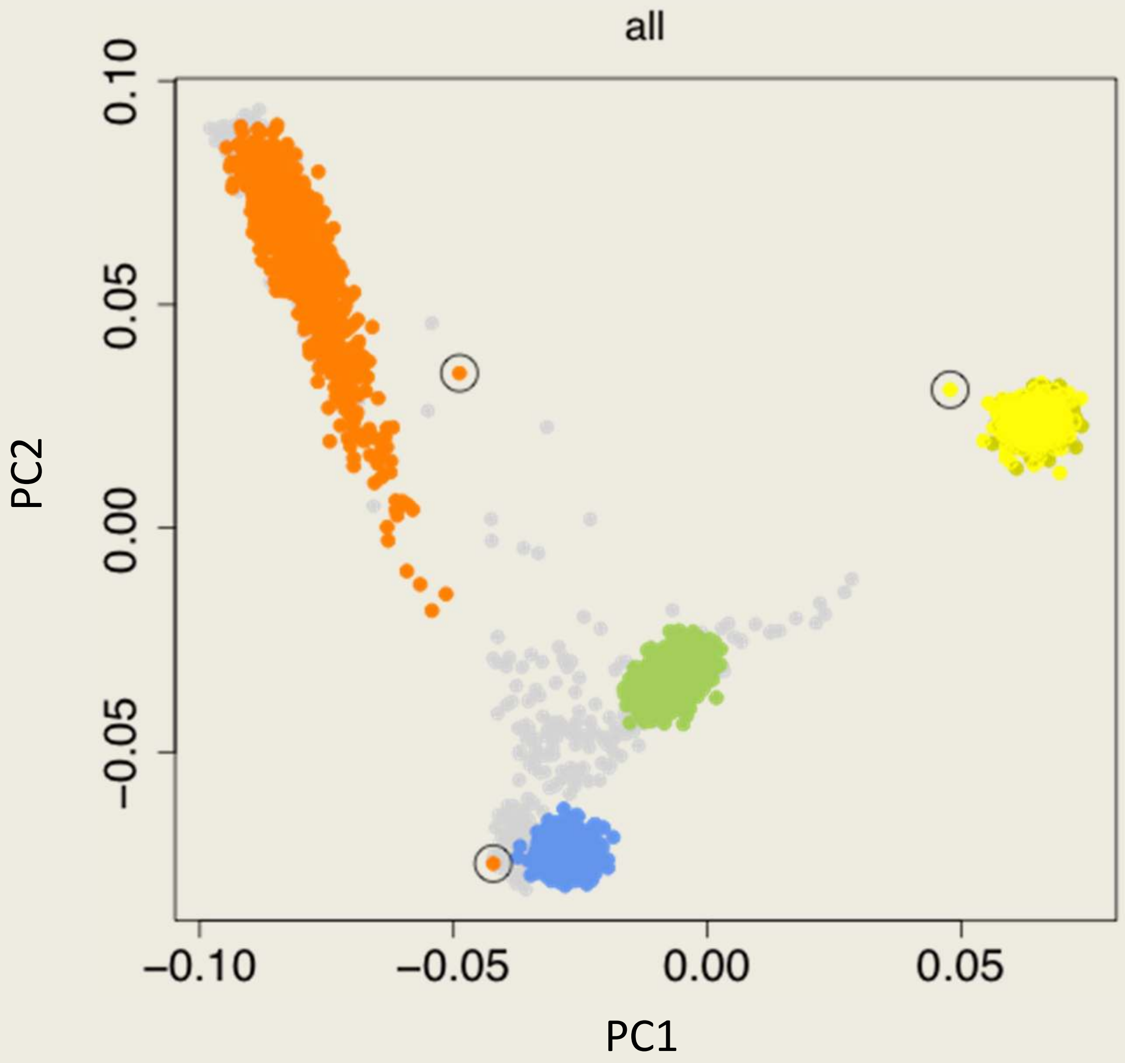
LOLIPOP* : the London Life Sciences Population Study

Samples & populations



- 10 cohorts(represent)
- 5 major ancestry groups

Samples & populations(cont.)



● African American

● East Asian (Korea)

● East Asian (Singapore)

● South Asian (Singapore)

● European (Finland)

Three outliers (circled) excluded from analysis

Samples & populations: Variant statistics

Table. SNP variation across cohorts (Autosomal only)
-- # Samples = 5334; # Variants = 1,768,095

Counts (%)	Wake Forest	KARE	Singapore Chinese	Singapore Indians	METSIM
# samples	1069	1093	1070	1140	962
# variants*	716,411	432,944	481,281	578,528	244,704
Private to EACH cohort					
# variants	490,155 (100)	219,813 (100)	255,203 (100)	366,525 (100)	78,208 (100)
# singletons	254,863 (52.0)	153,052 (69.6)	184,095 (72.1)	234,524 (64.0)	47269 (60.4)
# rare variants (< 1%)	425,467 (86.8)	219,064 (99.7)	254,493 (99.9)	357,805 (97.6)	74,501 (95.3)
# common variants (≥ 5%)	15,548 (3.2)	0 (0)	2 (0)	265 (0.07)	115 (0.15)
Shared across ALL cohorts					
# variants			71,062 (100)		
# singletons	1,336 (1.9)	2,431 (3.4)	2,251 (3.2)	1,204 (1.7)	1,784 (2.5)
# rare variants (< 1%)	8,995 (12.7)	9,913 (14.0)	10,088 (14.2)	6,969 (9.8)	8,652 (12.2)
# common variants (≥ 5%)	52,225 (73.5)	50,927 (71.7)	50,795 (71.5)	55,371 (77.9)	52,138 (73.4)

*Excludes 39,526 variants on chrX and 307 variants on chrY

Phenotypes

Variable	Column heading	Variable	Column heading
Diabetes disease status	T2D	Hip circumference	HIPC
Diabetes of diagnosis	AOD	Waist circumference	WAISTC
Fasting glucose	FAST_GLU	Diabetes medication	DIABMEDS
Fasting insulin	FAST_INS	Hypertension medication	BPMEDS
Fasting C-peptide	FAST_CPEP	Weight	WIEGHT
HbA1C	HBA1C	2-hour glucose	2H_GLU
GAD Ab	GAD	2-hour insulin	2H_INS
Creatine	CREATINE	2-hour C-peptide	2H_CPEP
Adiponectin	ADIPONECTIN	SEX	SEX
Leptin	LEPTIN	AGE	AGE
Total cholesterol	CHOL	Current use female hormone	HORMONES
LDL cholesterol	LDL	BMI	BMI
HDL cholesterol	HDL	Family ID	FAMID
Triglyceride	TG	STUDY ID	STUDYID
Height	HEIGHT	STUDY ID of father	FATHER
Systolic Blood pleasure	SBP	STUDY ID of mother	MOTHER
Diastolic Blood pleasure	DBP		

T2D Consortium

1. Introduction
2. Projects
   ① Project 1
   ② Project 2
   ③ Project 3
3. Our preliminary analysis results
   ① Single variant & Meta analysis
   ② Analysis of multiple variants

Project 2 : Introduction

- **Main task:** Detect rare (even private) functional variants influencing diabetes risk and diabetes-related phenotypes
- Assessed available pedigrees for potential to generate large number of copies of private variants, sequencing efficiency, diabetes prevalence
- Sequencing performed at Complete Genomics. ~600 samples at 60x coverage.

Project 2 : Introduction(cont.)

- Rare Variant Hypothesis
  - Human quantitative variation has a substantial component due to the effects of "rare" sequence variants in multiple genes.
  - Larger effects or rare variants will make disease related gene discovery easier.

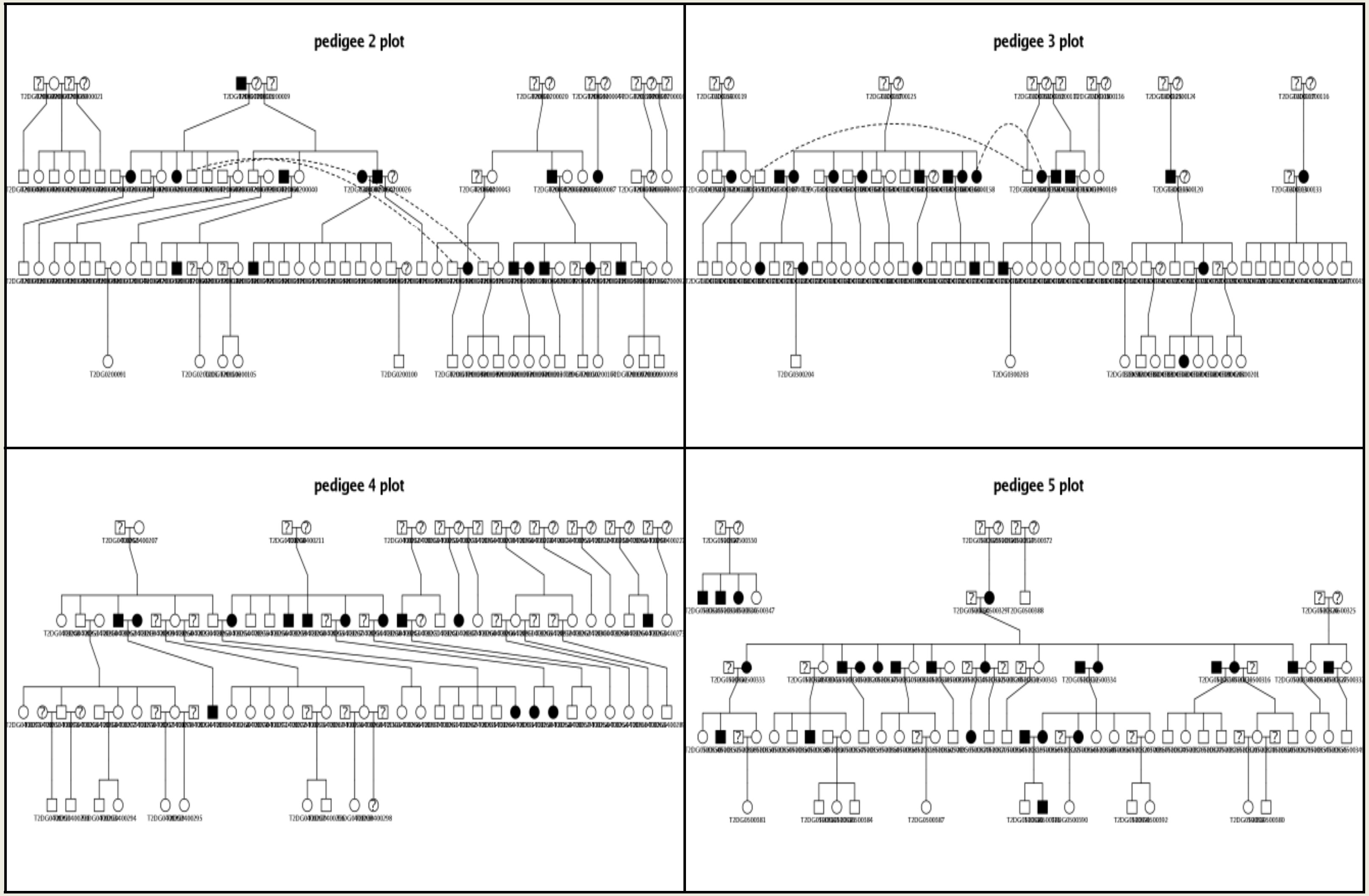
- How can we study Rare Variant
  - Very rare functional variants are best detected using a large pedigree based design.
  - Pedigrees allow observation of multiple copies of a private variant.

WGS in 20 Mexican American Pedigrees

- # of families : 20 families
- # of founders : 117 individuals

PEDIGREE	count
2	86
3	77
4	64
5	68
6	64
7	38
8	68
9	33
10	64
11	35
14	40
15	41
16	48
17	42
20	36
21	35
23	32
25	33
27	35
47	22

Mexican American Pedigrees plot



Phenotypes

- **Glycemic traits**
  - Fasting glucose
  - Fasting insulin
  - HbA1c
  - HOMA-B
  - HOMA-IR
- **Blood pressure**
  - SBP
  - DBP
- **Other biomarkers**
  - eGFR (creatinine)
  - Adiponectin
  - Leptin
  - GAD ab
- **Anthropometric traits**
  - Height
  - BMI
  - Waist circumference
  - Hip circumference
  - Waist to hip ratio
  - Lipids
  - HDL
  - LDL
  - Total cholesterol
  - triglycerides

T2D Consortium

1. Introduction
2. **Projects**
   ① Project 1
   ② Project 2
   ③ **Project 3**
3. Our preliminary analysis results
   ① Single variant & Meta analysis
   ② Analysis of multiple variants

Project 3 : Introduction

- Fine-mapping
  - ➢ Involves the identification of markers that are very tightly linked to a targeted gene.
  - ➢ Implies finding all the variants at the locus and trying to determine which changes may be related to pathogenesis with the use of statistical, functional, or bio-informatic methods.

Project 3 : Introduction(cont.)

- Meta-analysis of GWAS studies of T2D from diverse ethnic groups: European descent, South and East Asian descent, Hispanics and African-Americans.

- Initial focus on five T2D loci: CDKAL1, KCNQ1, CDKN2A/B, FTO and IGF2BP2 :
  - ➢ Strongest signals of association in most ethnic groups.
  - ➢ Evidence of differences in association signals and patterns of LD between ethnic groups.

Project 3 : Introduction(cont.)

- Summary of studies

Ethnic Groups	Study	Population	Ethnic Groups	Study	Population
European	WTCCC	UK	East Asian	HK1	Hong Kong
	FUSION	Finnish		HK2	Hong Kong
	LONGENETY	Askenazim		SGP-SIMES	Singapore Malay
	FHS	US		SGP-SP2	Singapore Chinese
	DGDG	French		CLHNS	Phillipino
South Asian	SGP-SINDI	Singapore Indian		JAPAN-KATO	Japanese
	PROMIS	Pakistani		JAPAN-KADO	Japanese
	INDIGO	North Indian	Mexican American	StarrCountry	Mexican American
	LOLIPOP	Indian	African American	JHS	African American

T2D Consortium

1. Introduction
2. Projects
   ① Project 1
   ② Project 2
   ③ Project 3
3. **Our preliminary analysis results**
   ① **Single variant & Meta analysis**
   ② Analysis of multiple variants

Project 1: KARE Data

- **GOAL**

  How do rare variants (MAF<0.05)Contribute to **T2D** and **BMI**?

- **KARE**(**K**orean **A**ssociation **RE**sources)
  - Exome data from 1093 Korean individuals
  - Independent sample : 1079
  - Related sample : 14

	MAF<0.01	0.01≤MAF≤0.05	MAF>0.05	Total SNP
Total	328,560 (74%)	24,320 (6%)	89,476 (20%)	442,356
Independent	326,377 (82%)	24,040 (6%)	49,312 (12%)	399,729

Methods

- Covariates
  - **T2D** : AGE+SEX+ AREA + AGE*SEX
  - **BMI** : AGE+SEX + AREA + AGE*SEX

- Methods

Phenotype	Independent Individual (1079)	Total individual (1093)
T2D	Logistic model	EMMAX (Kang et al, 2010 Nat Genet)
BMI	Linear model	

Project 1: Specific hypotheses

- Hypothesis 1
  - ➢ For any causal gene, the same rare variants will be associated (with similar effect) in all populations (Mega-analysis)
- Hypothesis 2
  - ➢ A causal gene will be associated with T2D in all populations, but with different causal variants and/or directions of effect (Meta-analysis)
- Hypothesis 3
  - ➢ Different causal genes will be associated in each population (Single-cohort analysis)

Methods for association analysis

Single-marker	EMMAX (Kang et al. Nat Genet. 2010)
	• Uses kinship to adjust for cryptic relatedness • Appropriately adjusts for population structure • 97% correlation with score test
Meta-analysis	MANTRA (Morris. Genet Epidemiol. 2011)
	• Assumes similar genetic effect for closely related populations, and heterogeneity between diverse groups

T2D : EMMAX Manhattan plot

MAF ≤ 0.01

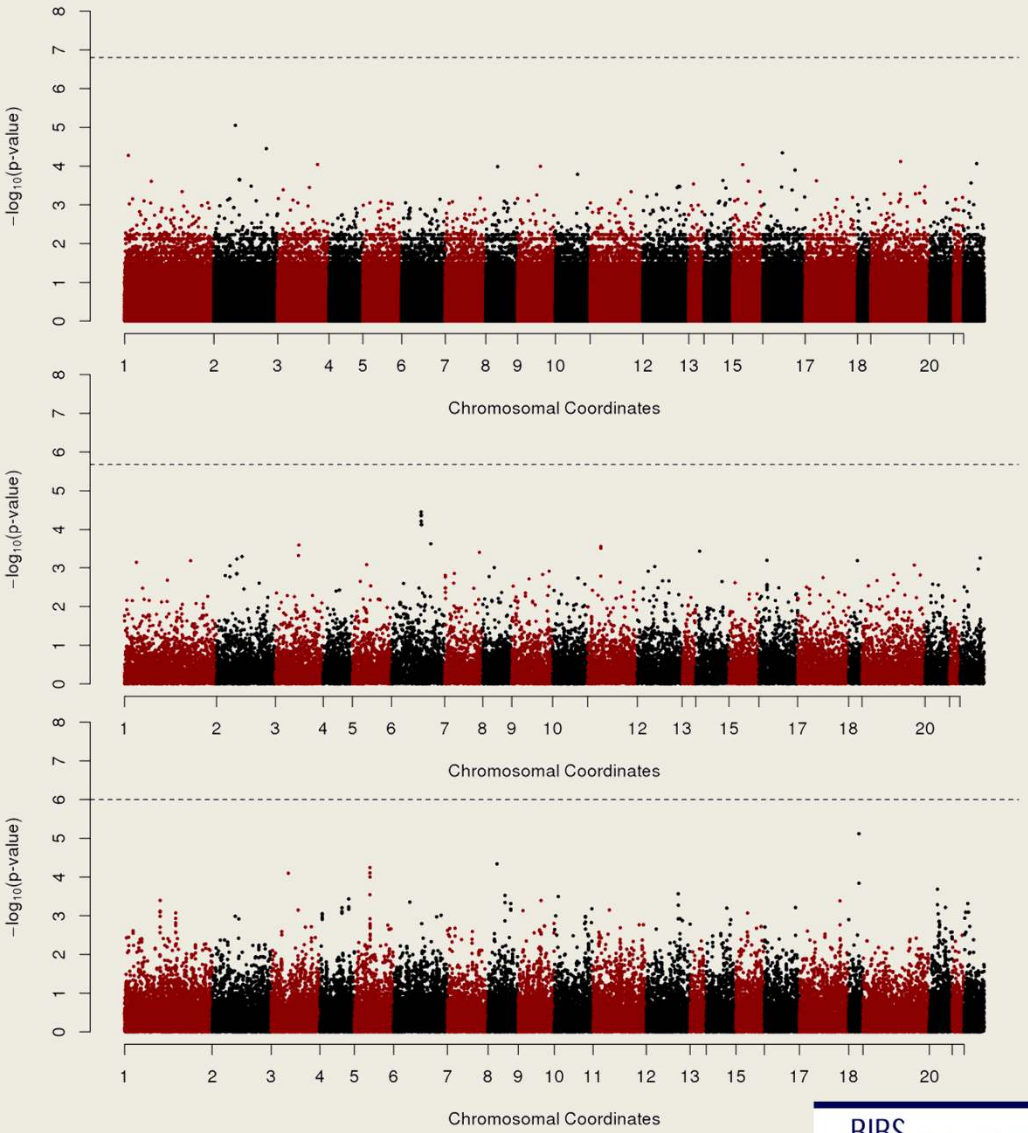
SNP = 328,560

0.01 < MAF ≤ 0.05

SNP = 24,320

MAF > 0.05

SNP = 89,476



T2D : Logistic regression Manhattan plot

MAF ≤ 0.01

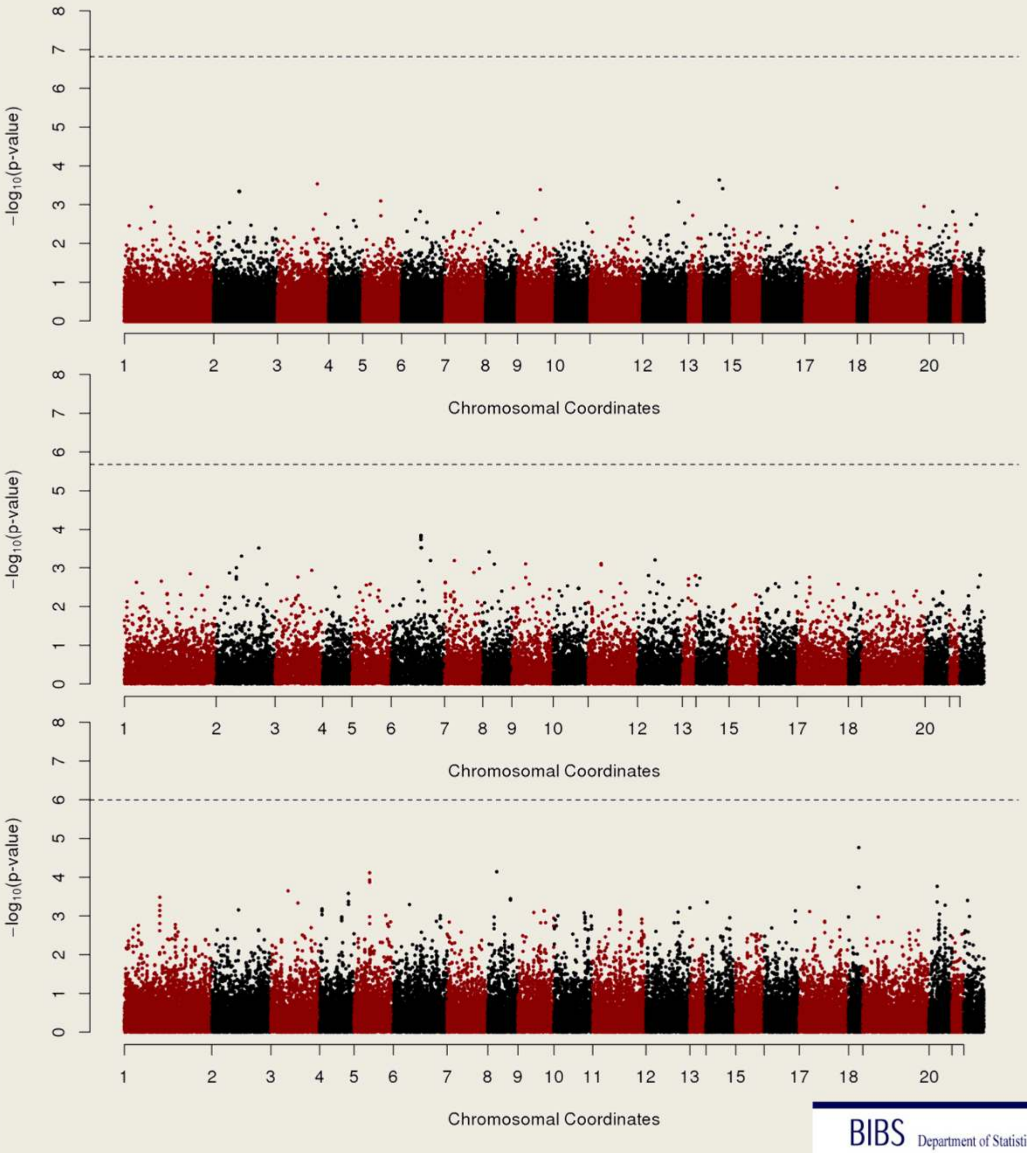
SNP = 326,377

0.01 < MAF ≤ 0.05

SNP = 24,040

MAF > 0.05

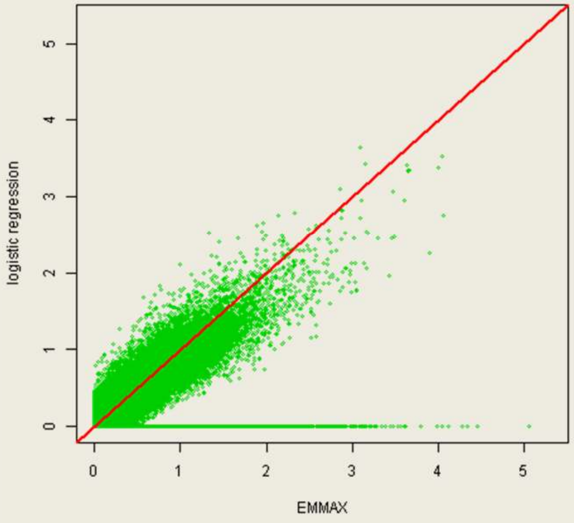
SNP = 49,312



T2D : EMMAX vs. logistic regression

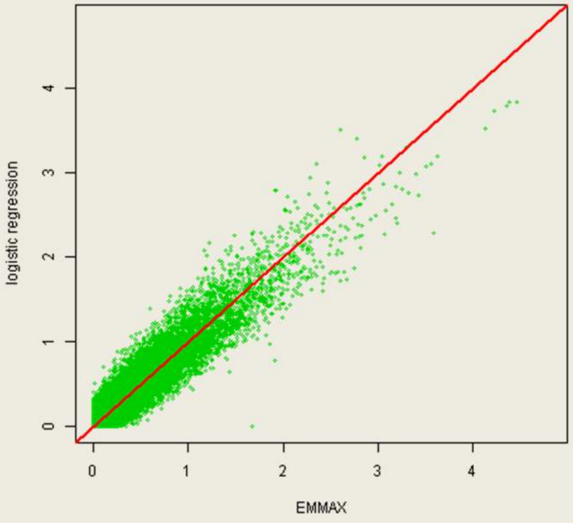
MAF≤0.01

$-\log_{10}$(p-value)



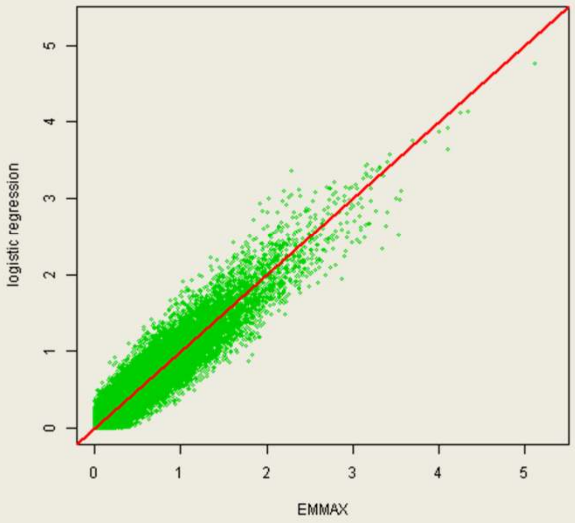
0.01<MAF≤0.05

$-\log_{10}$(p-value)



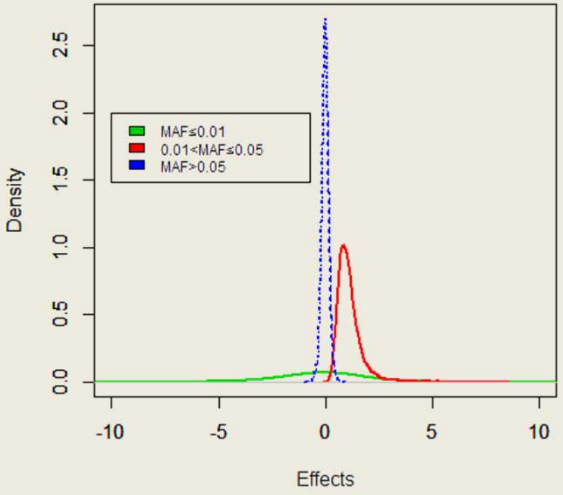
MAF>0.05

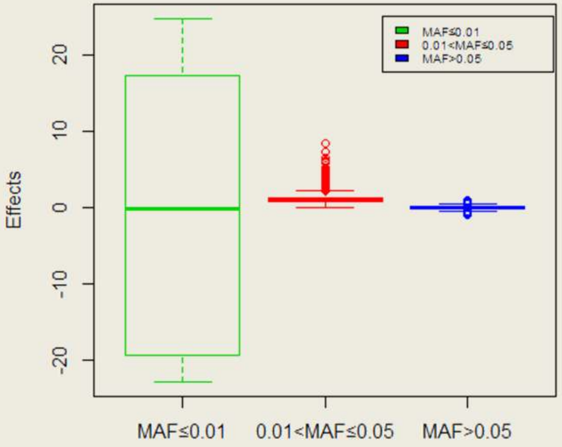
$-\log_{10}$(p-value)



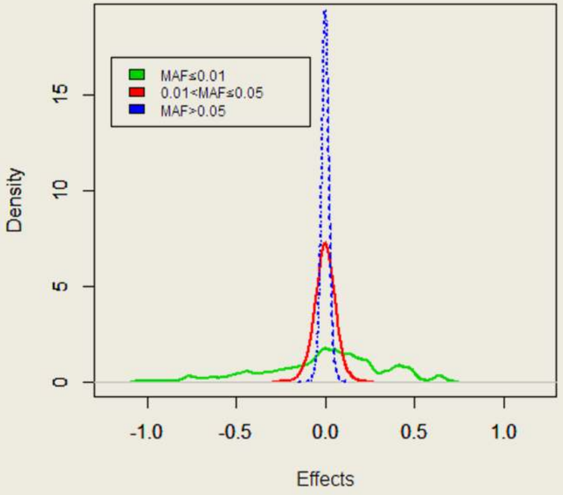
T2D : Effect size vs. MAF

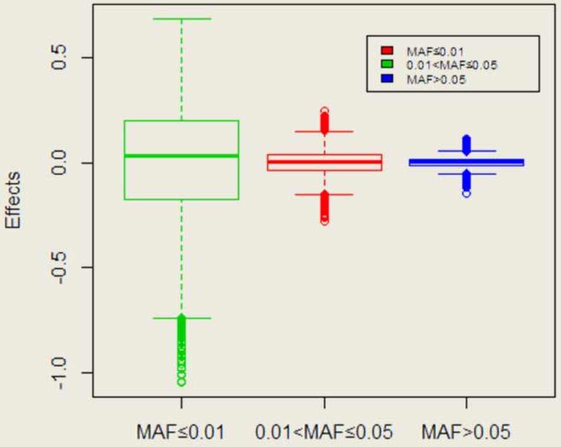
- Logistic Regression





- EMMAX





BMI : EMMAX Manhattan plot

MAF ≤ 0.01

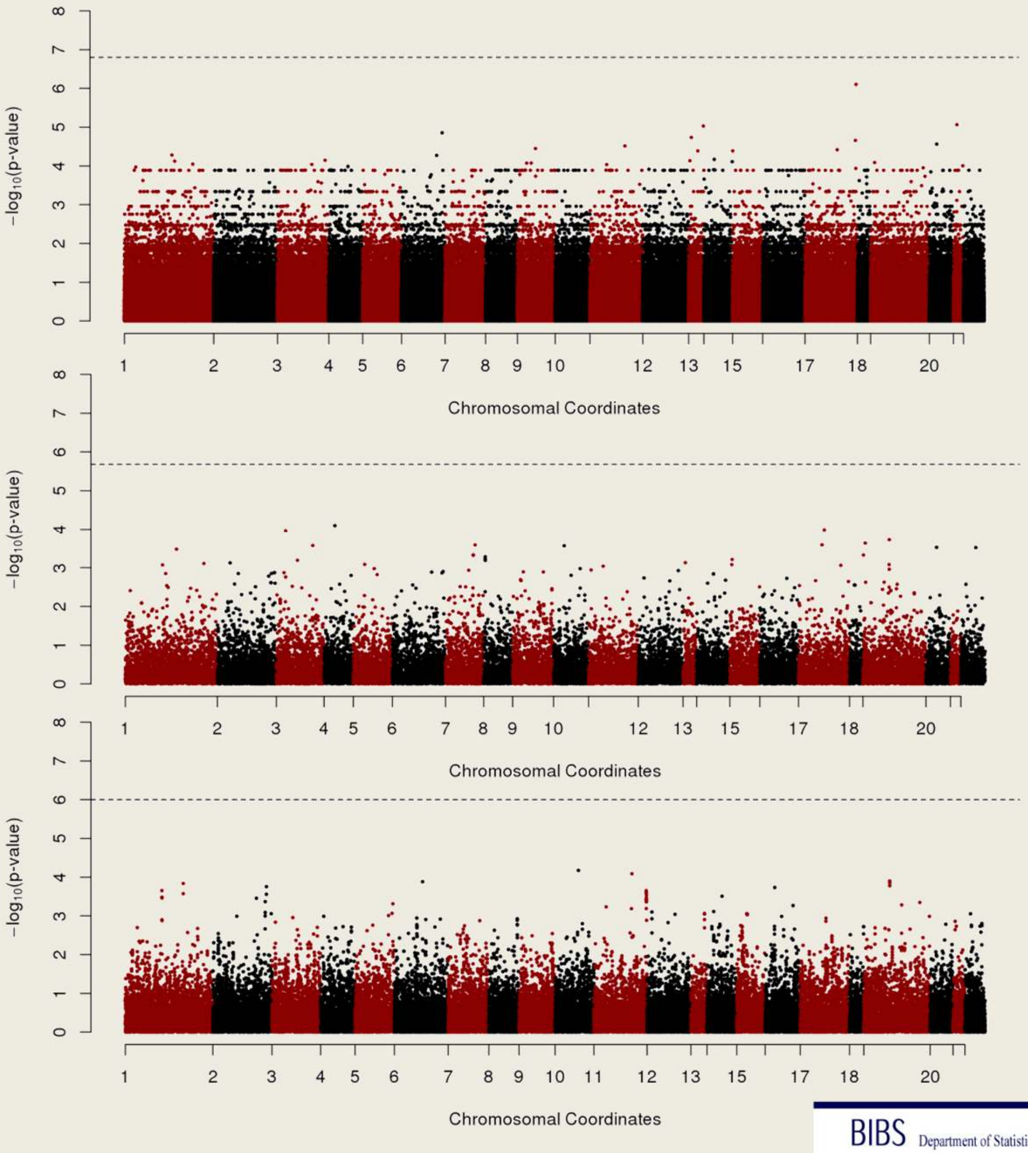
SNP = 328,560

0.01 < MAF ≤ 0.05

SNP = 24,320

MAF > 0.05

SNP = 89,476



BMI : Linear regression Manhattan plot

MAF ≤ 0.01

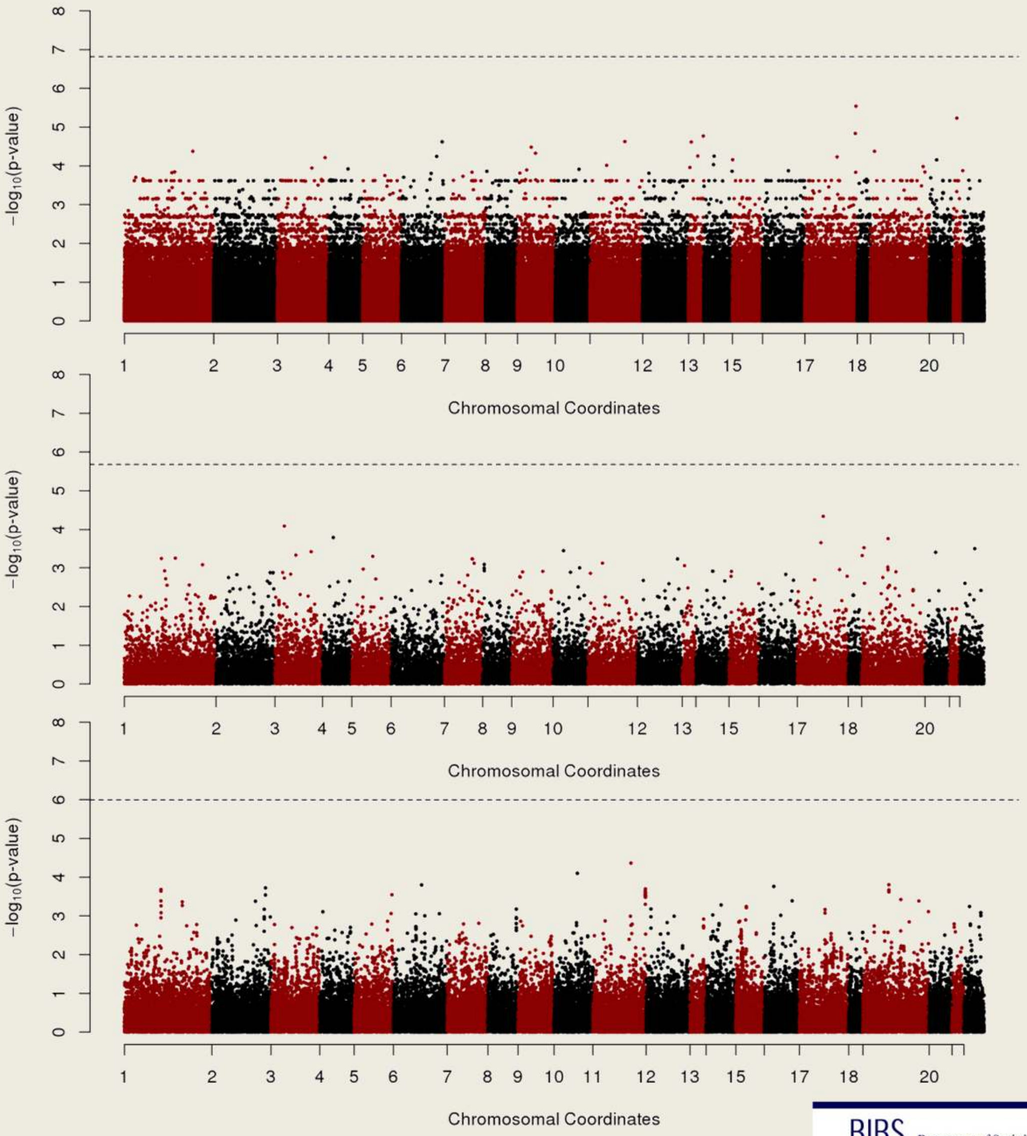
SNP = 326,377

0.01 < MAF ≤ 0.05

SNP = 24,040

MAF > 0.05

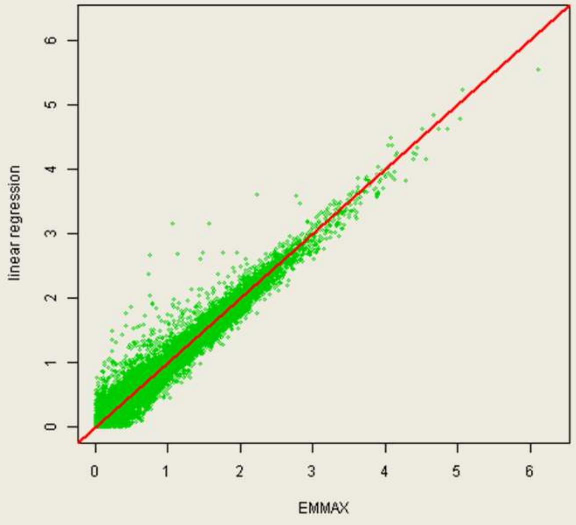
SNP = 49,312



BMI : EMMAX vs. linear regression

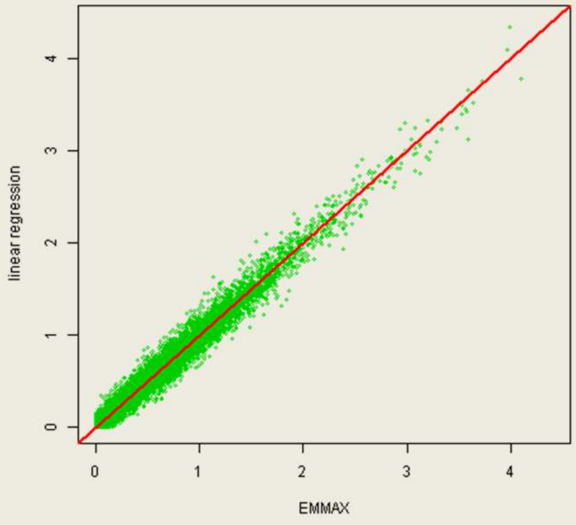
MAF≤0.01

-log$_{10}$(p-value)



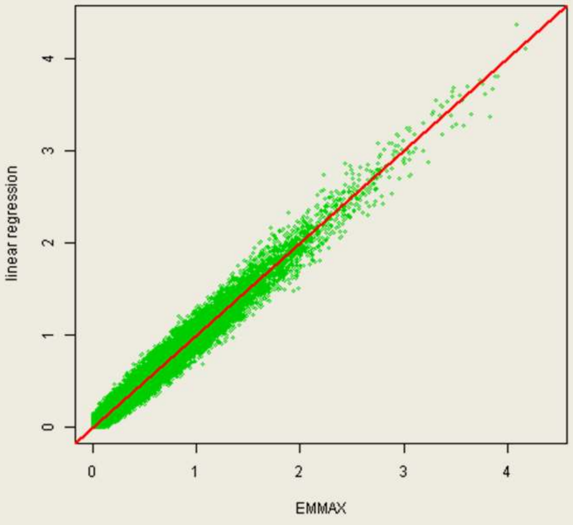
0.01<MAF≤0.05

-log$_{10}$(p-value)



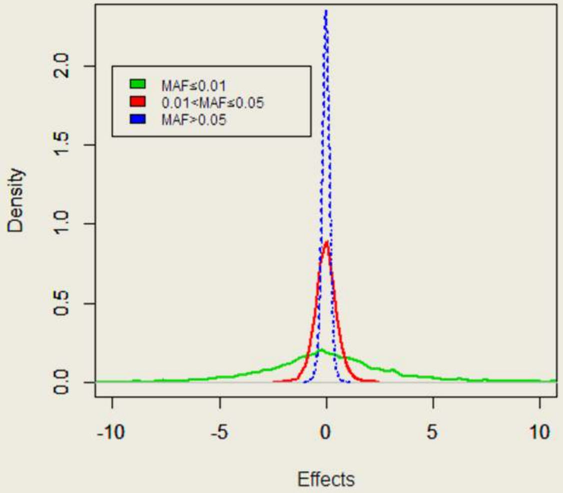
MAF>0.05

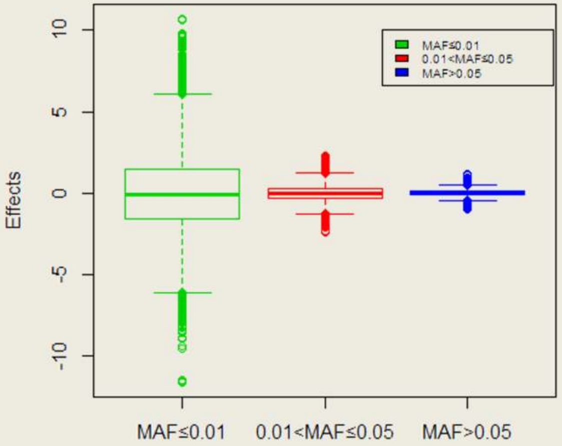
-log$_{10}$(p-value)



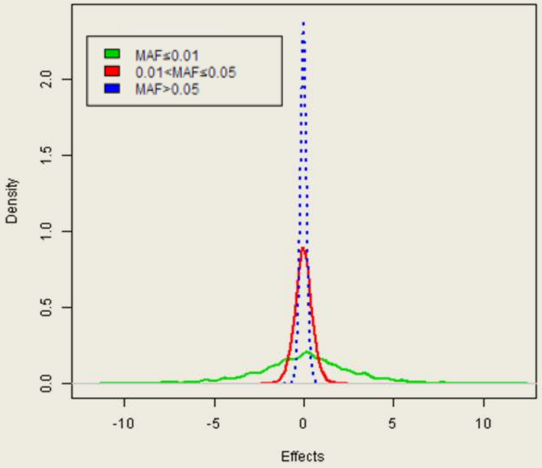
BMI : Effect size vs. MAF

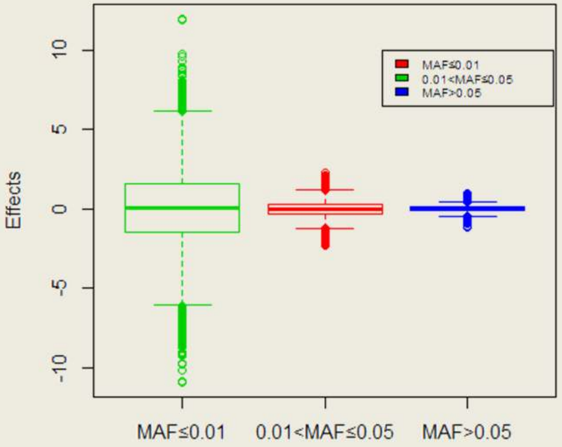
- Linear Regression





- EMMAX





Burden test for rare variants

- Since frequencies of rare variants are very low, even with high penetrance, it will be difficult to detect association with any single rare variants

- This has motivated **the development of new statistical tests** for detecting signals of rare variants

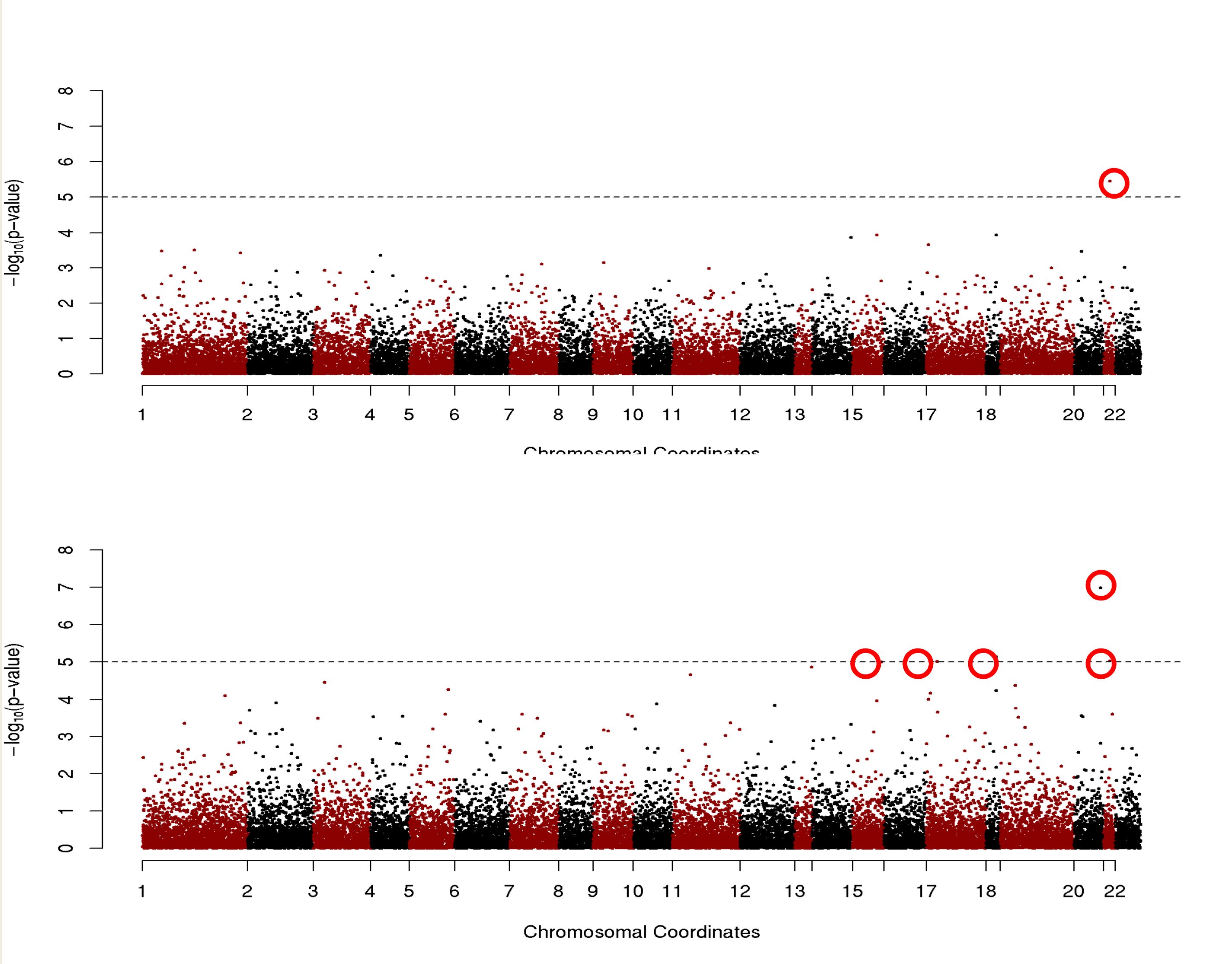
- Recent studies have shown that multiple rare variants could contribute to common diseases

- The burden test often employ the idea of **collapsing multiple rare variants** within a region

Burden tests for BMI

SKAT (SKAT-O)

Proposed test



sequence kernel association test

Improving powers in NGS Data

1. **Analysis of multiple rare variants**
   - ① Regularized Regression (Elastic-Net)
   - ② Gene-Gene Interaction

     Multifactor Dimensionality Reduction
   - ③ Gene Set Analysis
2. **Multivariate analysis**

Regularized regression for rare variants

- When the number of rare variants is large, the performance of burden test is not assured
- In this case, traditional regularized regression method (Ridge regression, LASSO, Elastic net, PCR and so on) could perform better than existing burden tests
- Each rare variant may not have stable association statistic especially when trait is binary
- Two step regularized regression method can be possible
  1. Collapse multiple rare variants into gene-level
  2. Conduct regularized regression for multiple gene-level variants

MDR for rare variants

- Interaction among rare variants do not occur as much as interaction among common variants

- If MDR is applied to rare variants, zero cell easily occurs and performance of MDR depends on one rare mutation very sensitively ➔ another version of MDR for rare variants should be developed

- If we collapse rare variants in gene-level first, then these collapsed variants could be used to conduct MDR

Gene set analysis for rare variants

- One-stage analysis (rare variants → Gene set)
  - Extension of gene-level burden tests (GRANVIL, SKAT, VT, WSS, and so on)
  - Traditional one-stage gene-set analysis
    - Sum of – log(p-value)
    - Enrichment score for chi-square statistics
- Two-stage analysis (rare variants → Gene → Gene set)
  - Traditional two-stage gene-set analysis
    - Highest chi-square + Enrichment score
    - Adaptive rank product +Adaptive rank product
    - Minimum p-value + network-based combined score
    - …

Gene set analysis for rare variants

- Limitations
  - If trait is binary, then association statistic (eg. p-value, chi-square statistic, …) for each rare variant is not stable
    - In this case, most of traditional gene set analysis cannot be applied
  - When a gene-set has a lot of variants (eg. ~1000 variants), then performance of burden tests are not assured yet
  - What if common variants and rare variants are together?
    - Most of burden test for rare variants give a larger weight to rarer variants
    - If burden tests are applied to real sequencing data, common causal variants will not be focused
    - Before combining rare variants and common variants, we should collapse multiple rare variants at first

Methods for Improving Powers

GWAS	Rare Variant Analysis
1. Single SNP analysis	1. Single SNP analysis Burden test
2. Meta analysis	2. Meta / Mega analysis
3. Analysis of multiple SNPs ① Regularized Regression ② Gene-Gene Interaction Multifactor Dimensionality Reduction ③ Gene Set Analysis	3. Analysis of multiple SNPs ① Regularized Regression ② Gene-Gene Interaction Multifactor Dimensionality Reduction ③ Gene Set Analysis
4. Multivariate analysis	4. Multivariate analysis

Acknowledgement

- **Bioinformatics and Biostatistics Lab., SNU**
  Sohee Oh, Dankyu Yoon, Min-Seok Kwon, Jaehoon Lee, Iksoo Huh, Seongyoung Lee
- **Center for Genome Science , KNIH, KCDC**
  Young Jin Kim, Joo-Young Lee
  Jong-Young Lee, Bok-Ghee Han,
- **KARE Consortium**
  Bermseok Oh, Kyunghee Univ.
  Sangoo Kim, Soongsil Univ.
- **KARE Cohort PIs**
  Nam Han Cho, Chol Shin
- **DNA-Link**
  Jong-Eun Lee
- **Sejong University**
  Seungyeoun Lee
- **University of Virginia**
  Ming Li
- **Case-Western University**
  Robert Elston

Acknowledgement

- T2D Project Consortium
  - **Broad Inst.**
    Nöel Burtt, Mark DePristo, Pierre Fontanillas
  - **Exter**
    Tim Frayling
  - **Harvard / MGH**
    Jose Florez, Jonna Grimsby, James Meigs
  - **National Univ. of Singapore**
    Yik Ying Teo, Xueling Sim
  - **Oxford**
    Mark McCarthy, Andrew Morris, Anubha Mahajan, Manny Rivas, Inga Prokopenko
  - **Singapore Eye Research Inst.**
    Kamran Ikram
  - **Univ. of Chicago**
    Nancy Cox, Mathew Barber
  - **Univ. of Michigan**
    Mike Boehnke, Tanya Teslovich, Christian Fuchsberger
  - **Univ. of Texas Health Sci. Ctr.**
    Craig Hanis, Taylor Maxwell, Heather Highland
  - **Wake Forest Univ. Sch. of Med.**
    Don Bowden, Maggie Ng
  - **Wellcome Trust Sanger Inst.**
    Ele Zeggini, Aaron Day-Williams

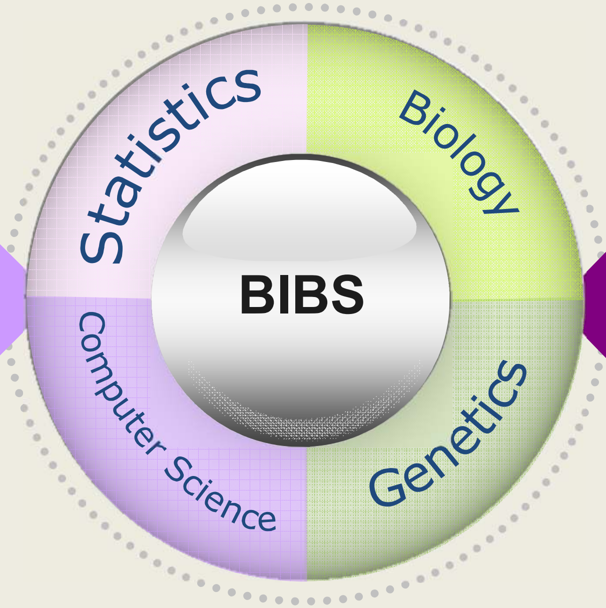
Acknowledgement
BIBS

Major: Biostatistics

- **Post-Doc:**
  Jiin Choi
  Shinik Kim

- **Ph. D. students :**
  Jaehoon Lee,
  Iksoo Huh,
- **M.S. student :**
  Seojin Bang,
  Yonggang Kim,
  Seungyeoun Hong
  Serong Lee,
  Byungju Min

IT + BT



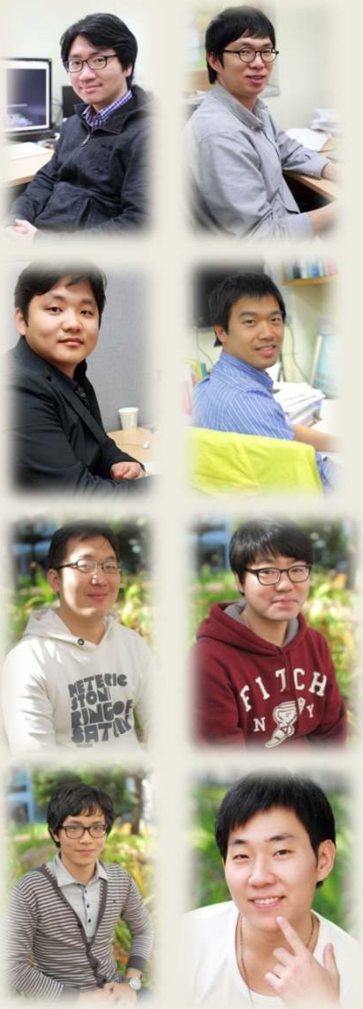
Major: Bioinformatics

- **Post-Doc:**
  Miae Doo

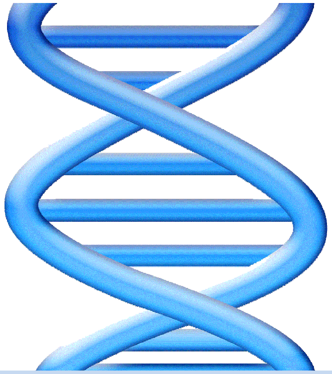
- **Ph. D. students :**
  Jungsoo Gim, Taejin Ahn, Minseok Kwon, Youngjin Kim, Sunkyoung Choi, Sungyoung Lee, Minseok Seo
- **M.S. students :**
  Sunghwan Cho,
  Eunyoung Ahn

창의연구단
National Creative
Research Initiatives

2012 BIBS members



Post docs



Thank you!

KARE: Type 2 Diabetes Characteristics

	Case	Control
# of Samples	1,042	2,943
Area (Ansung/Ansan)	531/511	1,669/1,274
Sex (Women/Men)	503/539	1,588/1,355
Age (Mean)	56.37	51.06
40th (%)	29.3	56.3
50th (%)	31.5	24.6
60 > (%)	39.2	19.1

- **Type 2 Diabetes**
  1) Treatment of T2D
  2) Fasting plasma glucose (FPG) ≥ 7 mmol/L or plasma glucose 2-h after ingestion of 75gm oral glucose load ≥ 11.1 mmol/L
  3) Age of disease onset ≥ 40 years

- **Controls**
  1) No history of diabetes
  2) FPG < 5.6 mmol/L and plasma glucose 2-h after ingestion of 75gm oral glucose load < 7.8 mmol/L at both baseline and follow up studies

Courtesy of KNIH

Three-Stage Approach for GWAS

- Stage I: Pre-screening for dimensionality reduction
  - Based on marginal regression
  - Selecting subset of SNPs showing strongest association with the trait
  - Sure Independence Screening (SIS, Fan & Lv, 2008)

- Stage II: Joint identification of putative causal SNPs via penalized regression with elastic net variable selection
  - Choice of optimal parameter λ
  - Based on10-fold cross validation
  - Minimizing prediction error rate

Three-Stage Approach for GWAS

Stage III: Validation of the jointly identified SNPs via EN based on Bootstrap Selection Stability(BSS)

- Investigate the consistency of the selected SNPs
- Use fixed optimal value of λ chosen at step II
- Elastic-net variable selection at each B=1000 bootstrap dataset

- Empirical replication of identified SNPs based on BSS is defined for *i* th SNP as follows:

$$BSS_i = \frac{1}{B}\sum_{b=1}^{B} I_i^b \quad , \text{ where } I_i^b = \begin{cases} 1 & \text{if replicated in } b^{th} \text{ bootstrap sample} \\ 0 & \text{otherwise} \end{cases}$$

Application of EN to KARE

Numbers of Height-Related SNPs

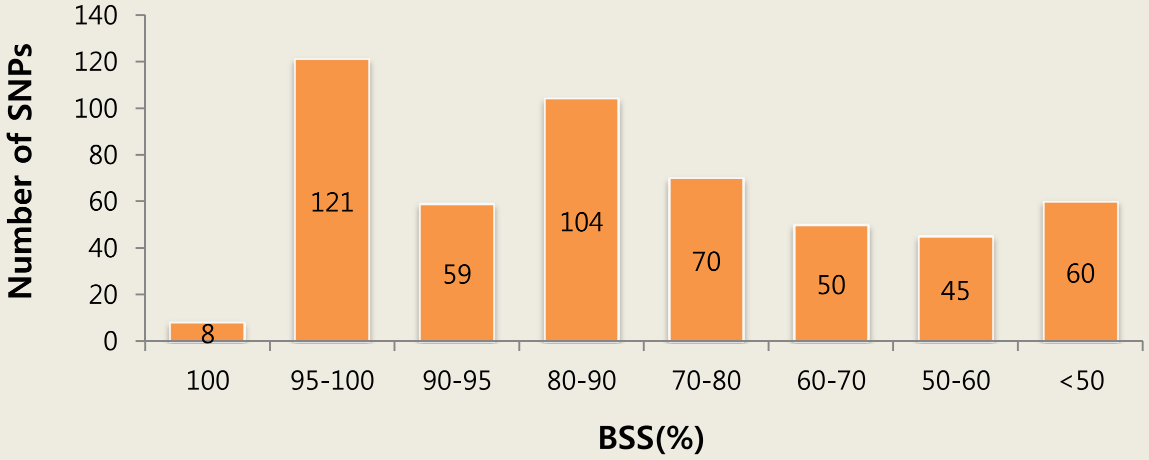
Step	# SNPs
Step 1 Screened SNPs	top1000
Step 2 Identified BMI-related SNPs	516 SNPs (208 known genes)
Step 3 Validation SNPs based on BSS	129 SNPs (64 known genes; BSS>95%)

Application of EN to KARE

Bootstrap selection stability(BSS)

- For each of 516 selected SNPs
  - 1000 bootstrap datasets with the fixed optimal λ
  - Compute BSS (%)

- Out of those 517 SNPs
  - 8 SNPs have 100% BSS
  - For 129 SNPs, BSS ≥ 95%
  - For 60 SNPs, BSS < 50%



Methods for Improving Power in GWAS

- <GWAS>

1. Single SNP analysis
2. Meta analysis
3. **Analysis of multiple SNPs**
   ① Regularized Regression
   ② Gene-Gene Interaction
   Multifactor Dimensionality Reduction
   ③ Gene Set Analysis
4. **Multivariate analysis**

- <Rare Variant Analysis>

1. Burden test?
2. Meta analysis
3. **Analysis of multiple SNPs**
   ① Regularized Regression
   ② Gene-Gene Interaction
   Multifactor Dimensionality Reduction
   ③ Gene Set Analysis
4. **Multivariate analysis**

MDR for rare variants

- Interaction among rare variants do not occur as much as interaction among common variants
- Interaction between rare variants and common variants should be considered
- If MDR is applied to rare variants, zero cell easily occurs and performance of MDR depends on one rare mutation very sensitively ➔ another version of MDR for rare variants should be developed
- If we collapse rare variants in gene-level first, then these collapsed variants could be used to conduct MDR

Gene set analysis for rare variants

- One-stage analysis (rare variants → Gene set)
  - Extension of gene-level burden tests (GRANVIL, SKAT, VT, WSS, and so on)
  - Traditional one-stage gene-set analysis
    - Sum of – log(p-value)
    - Enrichment score for chi-square statistics
- Two-stage analysis (rare variants → Gene → Gene set)
  - Traditional two-stage gene-set analysis
    - Highest chi-square + Enrichment score
    - Adaptive rank product +Adaptive rank product
    - Minimum p-value + network-based combined score
    - …

Gene set analysis for rare variants

- Limitations
  - If trait is binary, then association statistic (eg. p-value, chi-square statistic, …) for each rare variant is not stable
    - In this case, most of traditional gene set analysis cannot be applied
  - When a gene-set has a lot of variants (eg. ~1000 variants), then performance of burden tests are not assured yet
  - What if common variants and rare variants are together?
    - Most of burden test for rare variants give a larger weight to rarer variants
    - If burden tests are applied to real sequencing data, common causal variants will not be focused
    - Before combining rare variants and common variants, we should collapse multiple rare variants at first

Multivariate Analysis of KARE Data

- Different association direction in each phenotype
  - Multivariate has larger power than univariate analysis

CHR	SNP	BMI		Waist		Weight		WHR		Multivariate
		Beta	P-value	Beta	P-value	Beta	P-value	Beta	P-value	4 slopes
2	rs41498545	0.085	4.97E-02	-0.019	6.53E-01	0.044	2.56E-01	0.043	2.19E-01	1.38E-10
2	rs13410696	0.097	4.09E-02	-0.022	6.25E-01	0.119	4.50E-03	-0.012	7.56E-01	7.56E-11
3	rs17586294	0.038	2.55E-01	-0.076	1.97E-02	0.036	2.31E-01	-0.047	8.70E-02	3.86E-13
4	rs17501169	0.117	1.24E-02	-0.014	7.49E-01	0.096	2.00E-02	0.009	8.09E-01	3.06E-10
5	rs6866705	0.081	1.05E-01	-0.059	2.26E-01	0.088	4.94E-02	-0.023	5.71E-01	2.03E-11
6	rs6900453	0.065	1.28E-01	-0.056	1.79E-01	0.044	2.44E-01	-0.011	7.59E-01	2.87E-10
7	rs17168600	0.038	3.57E-01	-0.081	4.23E-02	0.026	4.72E-01	-0.035	3.01E-01	2.68E-10
11	rs17404578	0.004	9.29E-01	-0.111	4.65E-03	-0.003	9.29E-01	-0.037	2.66E-01	7.21E-13
11	rs41476549	0.112	1.26E-02	-0.007	8.66E-01	0.080	4.42E-02	0.042	2.53E-01	4.82E-12
18	rs11876341	-0.008	7.56E-01	0.018	4.60E-01	-0.033	1.39E-01	-0.020	3.36E-01	1.51E-10

Multivariate Analysis of KARE Data

- Same association directions

CHR	SNP	BP	BMI		Waist		Weight		WHR		Multivariate
			Beta	P-value	Beta	P-value	Beta	P-value	Beta	P-value	4 slopes
2	rs17584842	118474656	0.189	1.42E-02	0.095	2.03E-01	0.136	4.65E-02	0.182	3.82E-03	7.29E-08
2	rs1377819	125192366	-0.012	7.79E-01	-0.125	1.97E-03	-0.012	7.40E-01	-0.076	2.57E-02	1.31E-08
2	rs2360719	137105093	0.043	2.95E-02	0.001	9.71E-01	0.034	5.02E-02	0.014	3.76E-01	1.24E-07
3	rs6762722	142627906	0.055	1.50E-03	0.048	3.77E-03	0.080	1.59E-07	0.011	4.38E-01	1.23E-07
5	rs9327231	121131826	-0.006	9.01E-01	-0.134	5.07E-03	-0.002	9.69E-01	-0.089	2.72E-02	1.59E-07
7	rs4429999	70619015	0.199	1.47E-03	0.079	1.92E-01	0.116	3.72E-02	0.115	2.50E-02	4.98E-07
7	rs7792191	120165184	-0.001	9.89E-01	-0.140	1.12E-02	-0.020	6.86E-01	-0.071	1.27E-01	7.54E-07
10	rs2804219	117303461	0.064	3.86E-01	0.194	6.23E-03	0.030	6.40E-01	0.086	1.53E-01	5.24E-07
11	rs17145229	82862073	0.106	3.96E-02	0.002	9.67E-01	0.095	3.73E-02	0.046	2.75E-01	1.04E-08
12	rs1371090	89107773	0.159	8.78E-04	0.015	7.40E-01	0.118	5.45E-03	0.009	8.08E-01	7.51E-09
14	rs17109739	79218034	-0.003	9.42E-01	-0.091	2.08E-02	-0.012	7.33E-01	-0.032	3.35E-01	4.31E-07
16	rs16951883	10226280	0.115	4.87E-02	0.145	1.02E-02	0.111	3.09E-02	0.233	9.33E-07	2.35E-09

MDR: Overview

- Step1.
  - Identify the best combination of factors like SNPs and discrete environmental factors

- Step 2.
  - Define levels that are associated with the high risk of disease and levels that are associated with low risk

MDR: Overview

(SNP1, SNP2)	# of cases	# of controls	#case/#cont
(AA, BB)	50	40	1.25 High
(AA, Bb)	30 TP	25 FP	1.20 High
(AA, bb)	20	30	0.67 Low
(Aa, BB)	40 FN	45 TN	0.89 Low
(Aa, Bb)	25	30	0.83 Low
(Aa, bb)	20	10	2.00 High
(aa, BB)	10	18	0.56 Low
(aa, Bb)	3	1	3.00 High
(aa, bb)	2	1	2.00 High
Total	200	200	

High Risk Group

$$\Leftrightarrow \frac{n_{ij}^{\text{case}}}{n_{ij}^{\text{ctl}}} \geq \frac{n^{\text{case}}}{n^{\text{ctl}}}$$

Low Risk Group

$$\Leftrightarrow \frac{n_{ij}^{\text{case}}}{n_{ij}^{\text{ctl}}} < \frac{n^{\text{case}}}{n^{\text{ctl}}}$$

		Disease	
		Case	Control
Risk	High	105	77
	Low	95	123

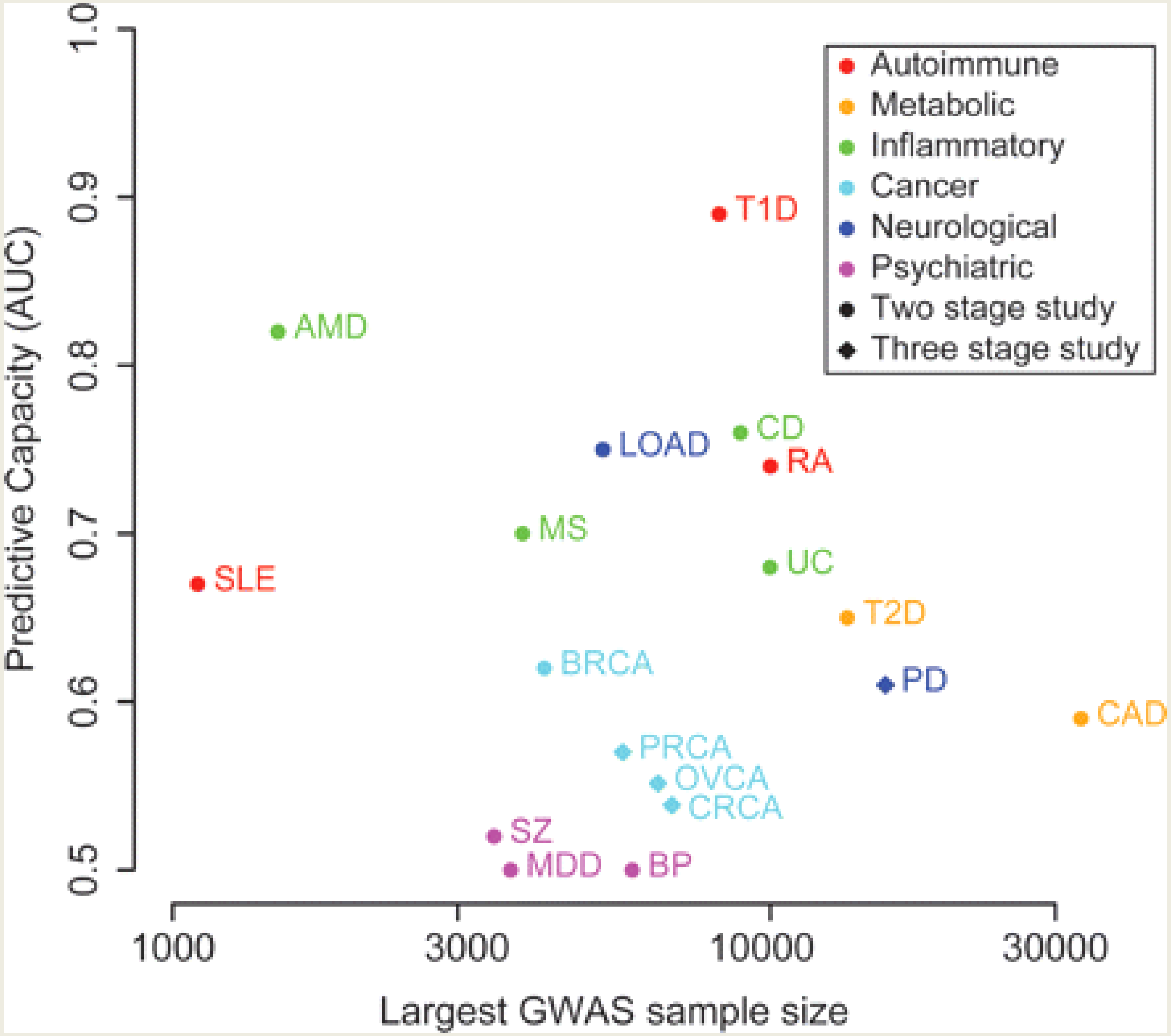
MDR: Overview

Set of SNPs : {SNP1, SNP2, … , SNP10}

	Two-dimensional	Three-dimensional	Four-dimensional
CV1	(SNP2, SNP6)	(SNP2, SNP5, SNP10)	(SNP2, SNP5, SNP6, SNP9)
CV2	(SNP4, SNP5)	(SNP1, SNP6, SNP10)	(SNP2, SNP6, SNP7, SNP10)
	………	………….	……………
CV10	(SNP1, SNP6)	(SNP2, SNP5, SNP10)	(SNP2, SNP5, SNP6, SNP9)
	⬇	⬇	⬇
	(SNP2, SNP6)	(SNP2, SNP5, SNP10)	(SNP2, SNP5, SNP6, SNP9)

SNPs	Balanced Accuracy	CV Consistency
(SNP2, SNP6)	0.75	9.0
(SNP2, SNP5, SNP10)	0.65	5.1
(SNP2, SNP5, SNP6, SNP9)	0.53	7.2

Type 2 Diabetes Risk Prediction



Jostins and Barrett *Human Molecular Genetics 2011*

Acknowledgement

- **Center for Genome Science, KNIH, KCDC**
  Young Jin Kim, Jong-Young Lee
  Bok-Ghee Han,
- **KARE Consortium**
- **Ansung and Ansan Cohort PIs**
  Nam Han Cho, Chol Shin
- **DNA-Link**
  Jong-Eun Lee
- **T2D Consortium PIs**
  Mike Boehnke, Mark McCarthy, David Altshuler, John Biangero, Nancy Cox
- **University of Virginia**
  Ming Li
- **Case-Western University**
  Robert Elston