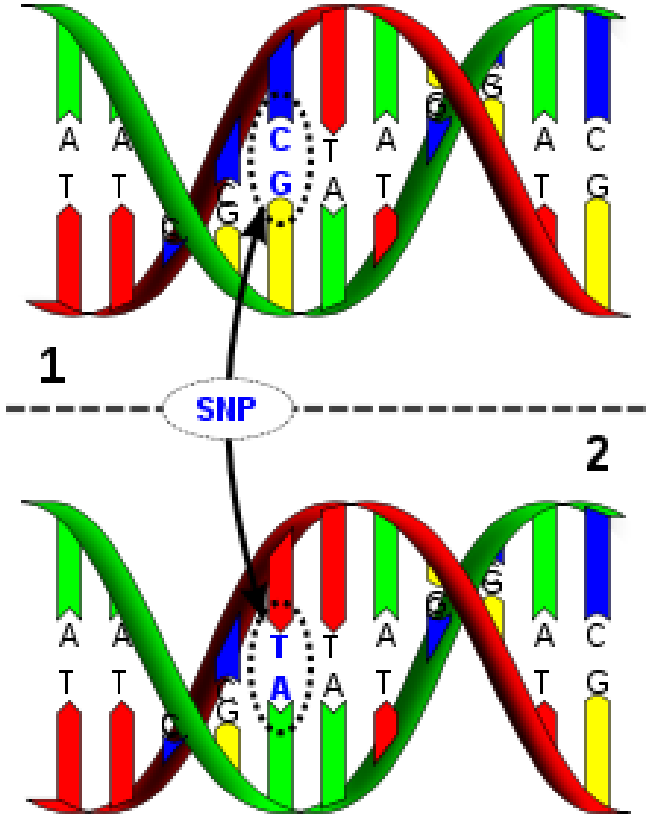
How to increase power of GWAS: pathway and meta analysis

UNIST
Dougu Nam

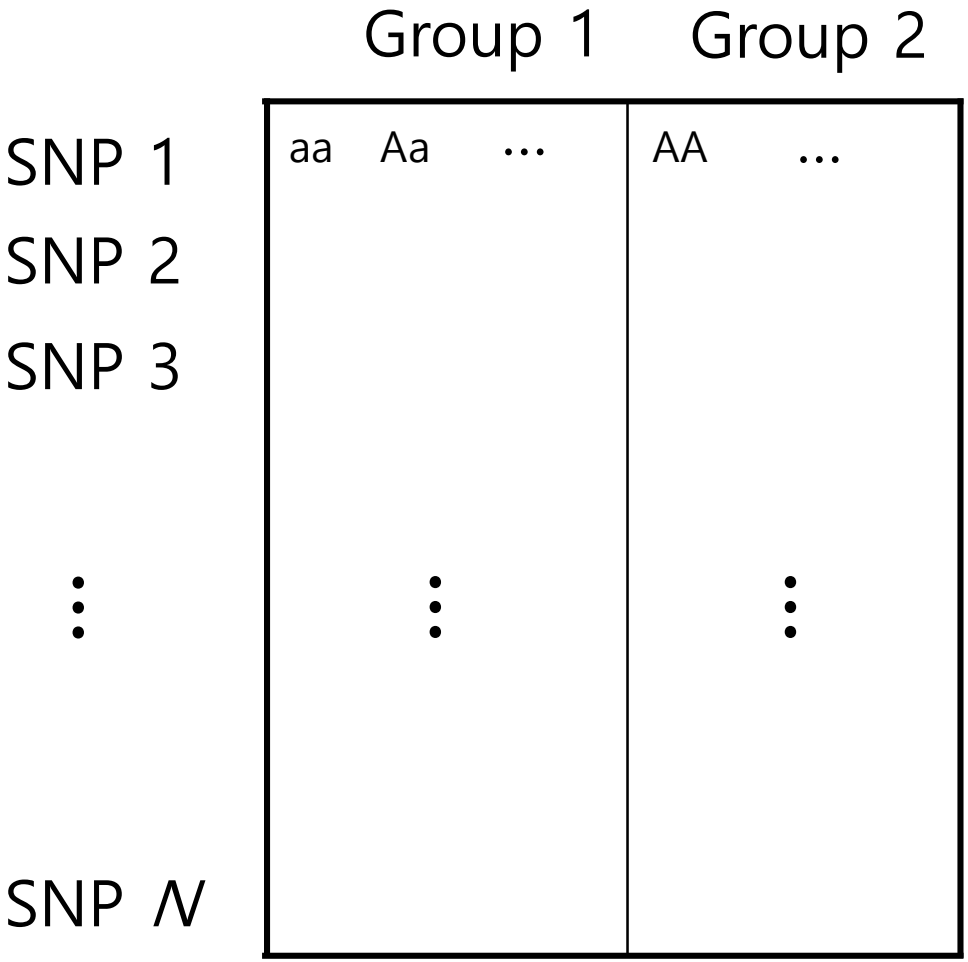
Single Nucleotide Polymorphism (SNP)



- Wikipedia

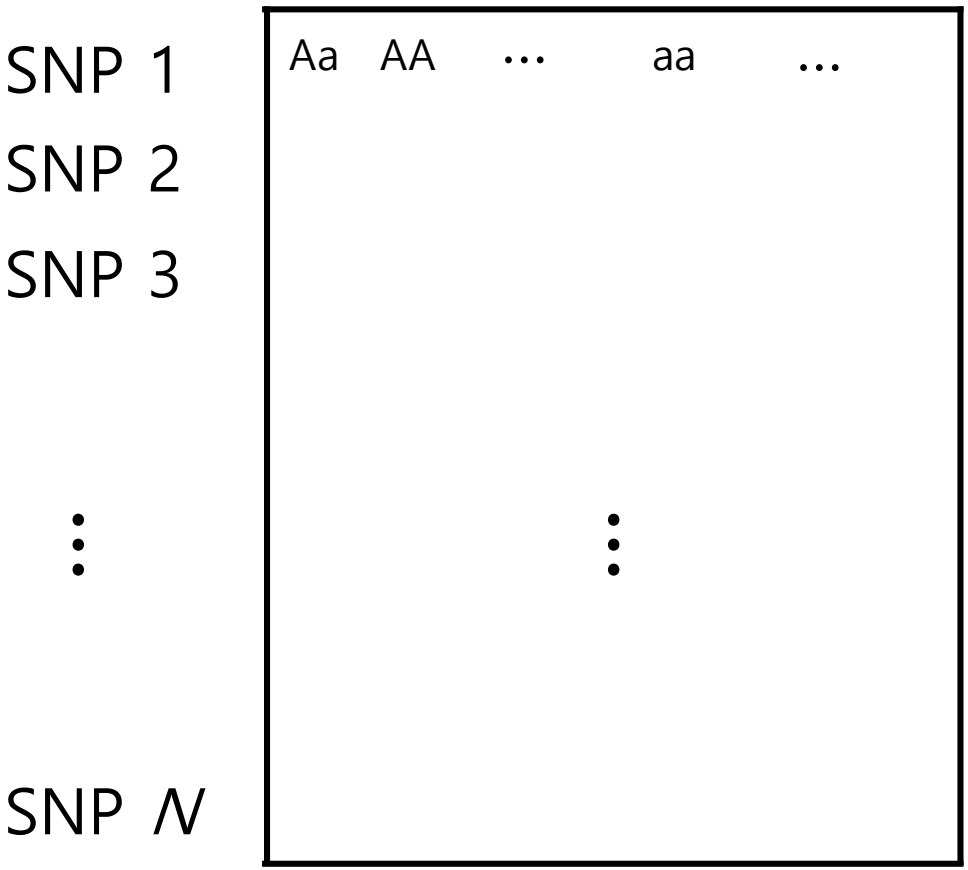
- A variation at a single site in DNA, is the most frequent type of variation in the genome (~ over 10 million)
- Responsible to most genetic disease and other phenotypes
- **Genome-wide association study** (**GWAS**) seeks to find relevant SNPs that may cause specific disease

GWAS: **Case-Control** study



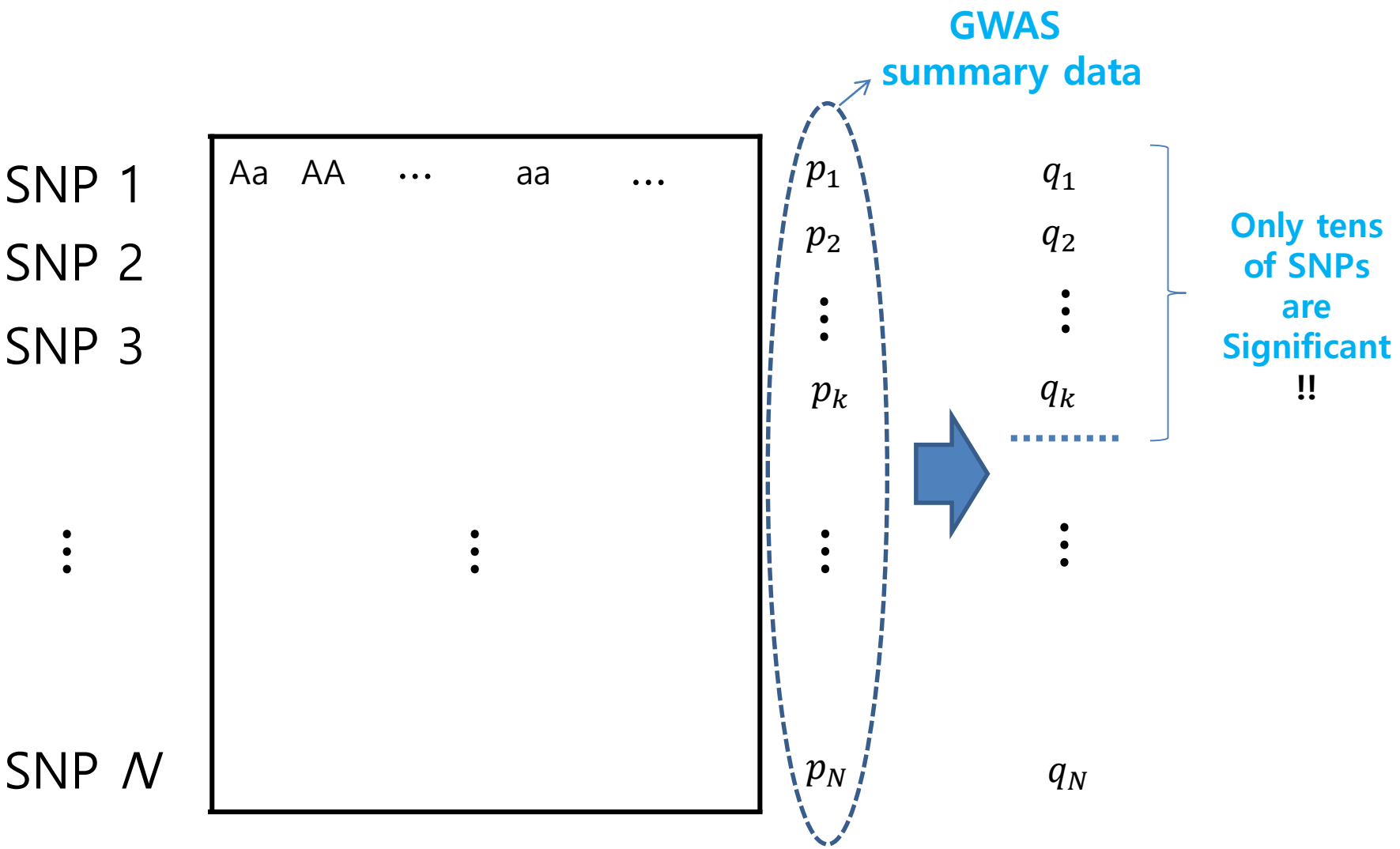
- Unlike gene expression data, SNP array provides **categorical data** (AA, Aa, or aa)
- We are interested in identifying <u>which SNP is associated with a given disease</u>

GWAS: **Quantitative Trait** study



- **<u>No partition</u>** on samples
- Each sample has some <u>continuous phenotype values</u> such as height, blood pressure

GWAS

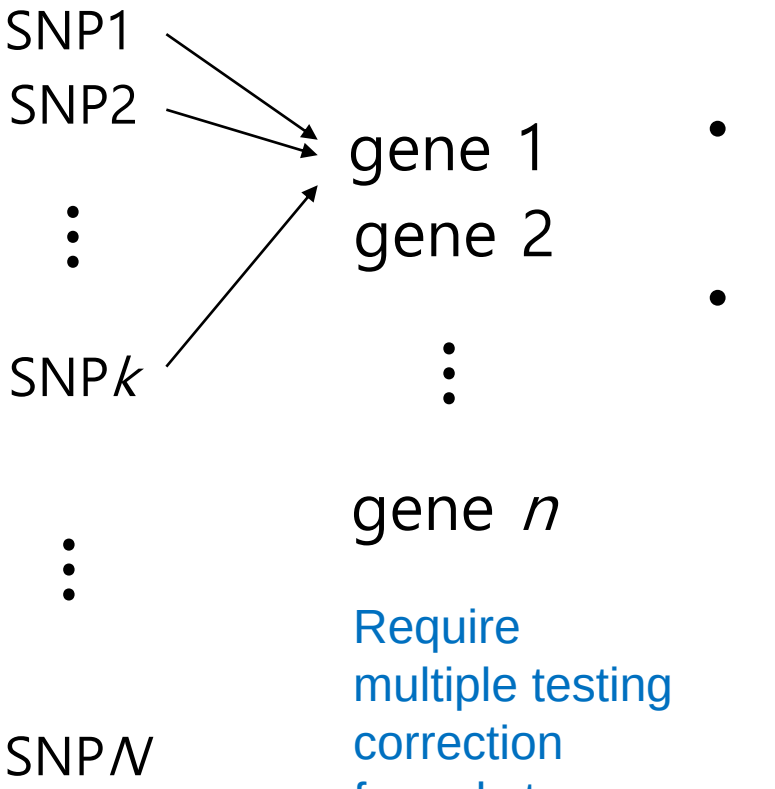

GWAS
summary data
SNP 1
SNP 2
SNP 3
SNP N
Aa
AA
…
aa
…
p_1
p_2
p_k
p_N
q_1
q_2
q_k
q_N
Only tens
of SNPs
are
Significant
!!

General Problems in GWAS

- Needs multiple testing correction for a million of *p*-values
- A very stringent cutoff (e.g. p=10^-8) is used to yield only a small number of SNPs
- Many moderate but meaningful associations outside the cutoff is lost. This is **very inefficient and wasteful**
- It is not easy to discuss biology only with a small number of significant SNPs

Tip of Iceberg

Increasing power of GWAS: gene-based test

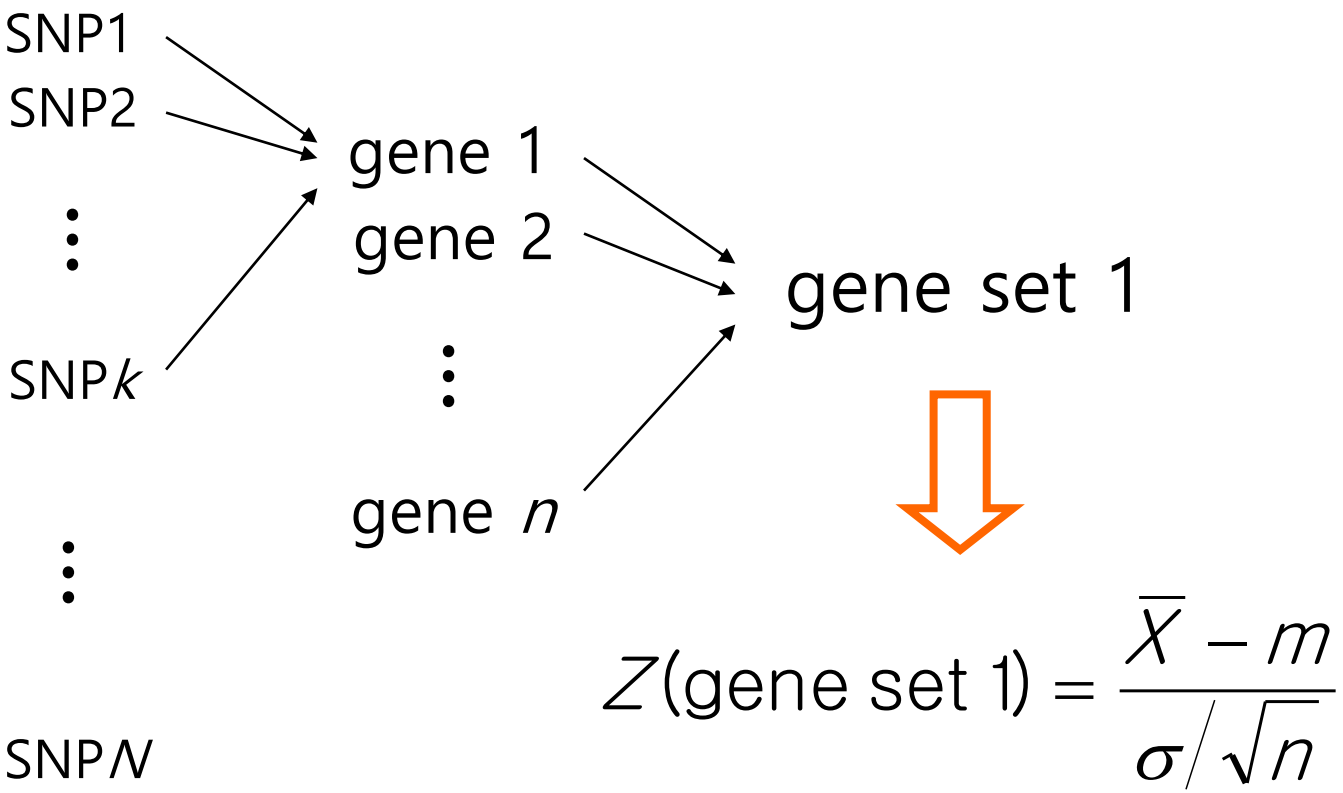


- Summarize SNPs to genes
- **VEGAS method**: Given *k* SNP *p*-values for a gene, correlated *p*-values are simulated using multivariate normal distribution, and the summarized statistic values are assessed (Liu et al. AJHG 2010)

Increasing power of GWAS: gene-set-based test

- Pathway (gene-set) analysis was also considered for GWAS data to find missing heritability
- This approach aims to detect **moderate but coordinated** associations within a gene set (as well as strong signals)
- **GSEA** was firstly introduced for GWAS genotype data (Wang et al. AJHG 2007): requires heavy calculation

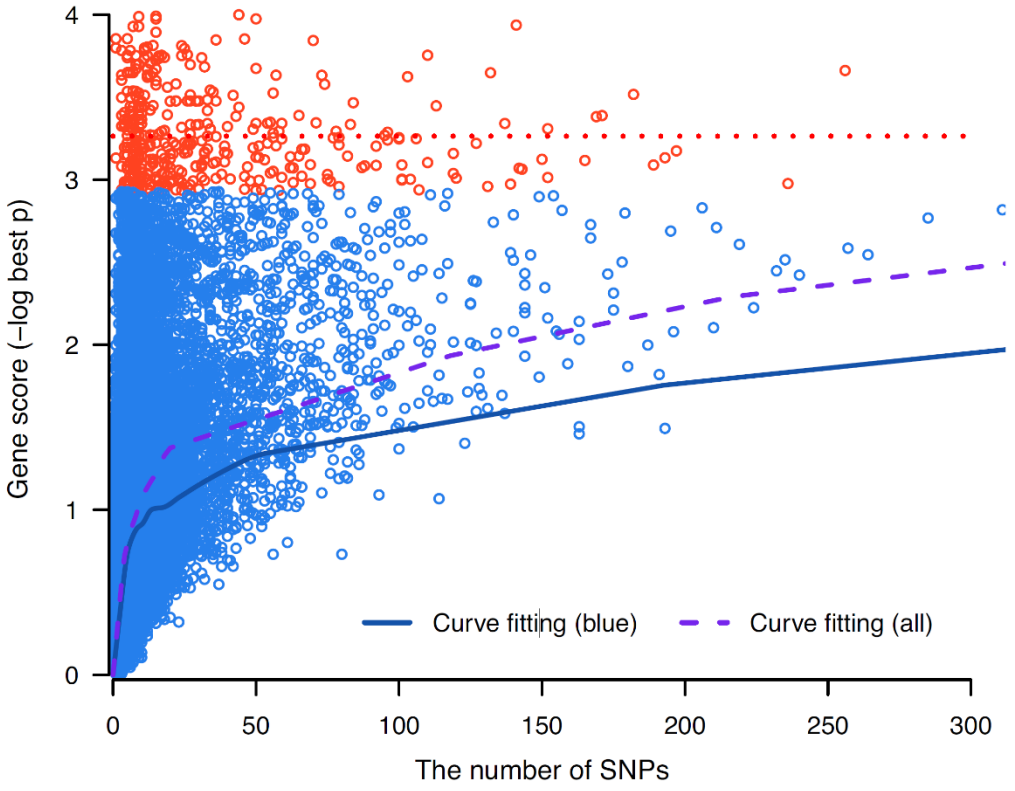
Increasing power of GWAS: gene-set-based test



GSA-SNP (Nam et al. 2010 NAR):
Uses only summary *p*-values
to calculate pathway statistic

How to improve the method?: GSA-SNP2 (Yoon et al. 2018 NAR)

- SNP size effect
- SNP-SNP correlation adjustment



- Monotone cubic spline

$$Adj(g_i) = -\log(\text{best } p_i) - \mathrm{C}(g_i)$$

- Pathway score:

$$Z\left(P_j\right) = \frac{\overline{P_j} - m}{\sigma/\sqrt{N_j}}$$

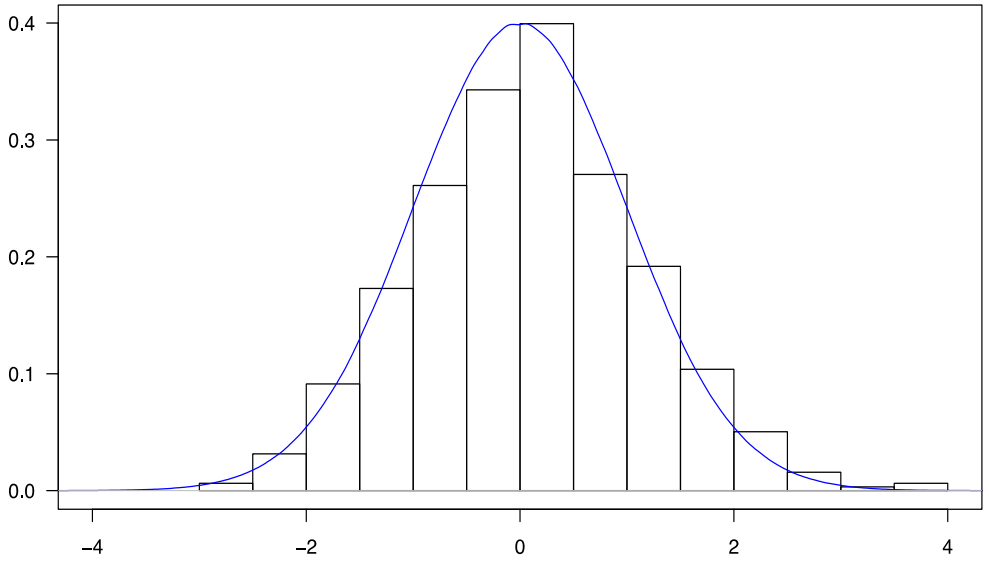
Existing methods

- MAGENTA (PLoS Genetics, 2010):
  - Adjust for confounding factors using regression model

$$Z_g^{BestSNP} = \alpha \cdot d_g + \beta \cdot n_g + \delta \cdot u_g + \gamma \cdot h_g + \eta \cdot c_g + \kappa \cdot l_g + r_g$$

  - Strict false positive control
  - Very low power

Distribution of pathway Z-scores



- Distribution of the 674 Reactome pathway Z-statistic of adjusted gene scores for GWAS data simulated using 1000 Genomes data. Standard normal distribution is fitted (blue curve)
- **This implies we don’t have to deal with the heavy genotype data for pathway analysis**

Simulation test

- Used real genotype data: 1000 Genomes
- Simulated phenotypes using linear model
  - False positive control:

$$Y = \beta_1 X_1 + \cdots \beta_k X_k + \varepsilon$$

  - Statistical Power

$$Y = \beta_1 X_1 + \cdots \beta_k X_k + \gamma\,(G_1 + \cdots + G_M) + \varepsilon$$

Simulation test

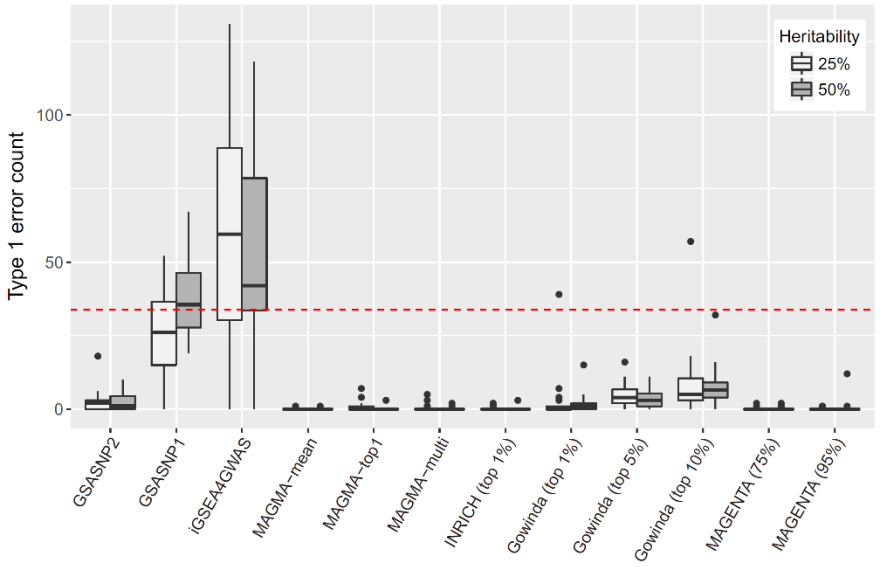
- Background heritability:

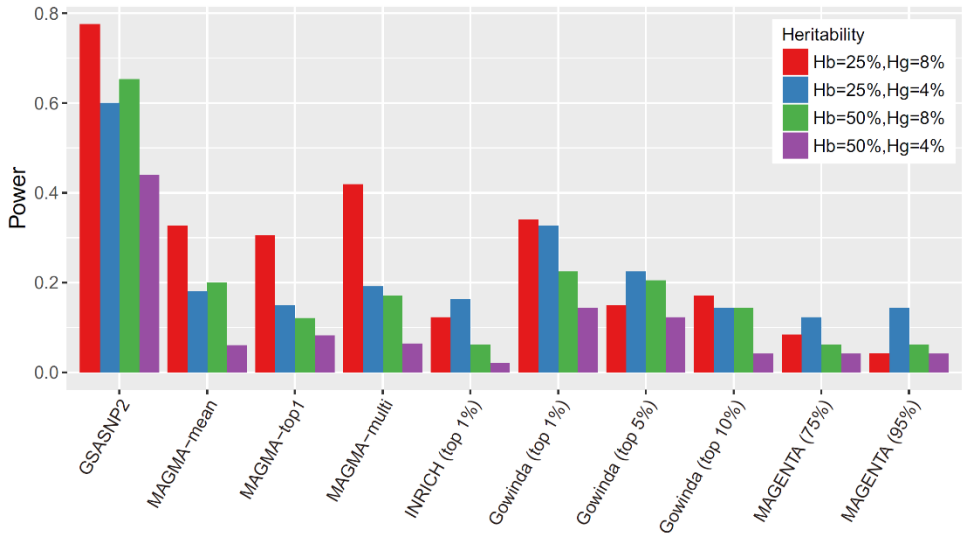
$$h_b^2 = \frac{Var(\beta_1 X_1 + \cdots \beta_k X_k)}{Var(Y)}$$

- Gene-set heritability:

$$h_g^2 = \frac{Var(\gamma(G_1 + \cdots + G_M))}{Var(Y)}$$

- Used only 10,000 samples





Tests for real GWAS data

- Used publicly available summary data from DIAGRAM and GIANT consortia
- Gold standard pathways for T2D: collected by third party, Morris et al.
- Gold standard pathways for human height: collected from the literature

T2D data analysis results by GSA-SNP2

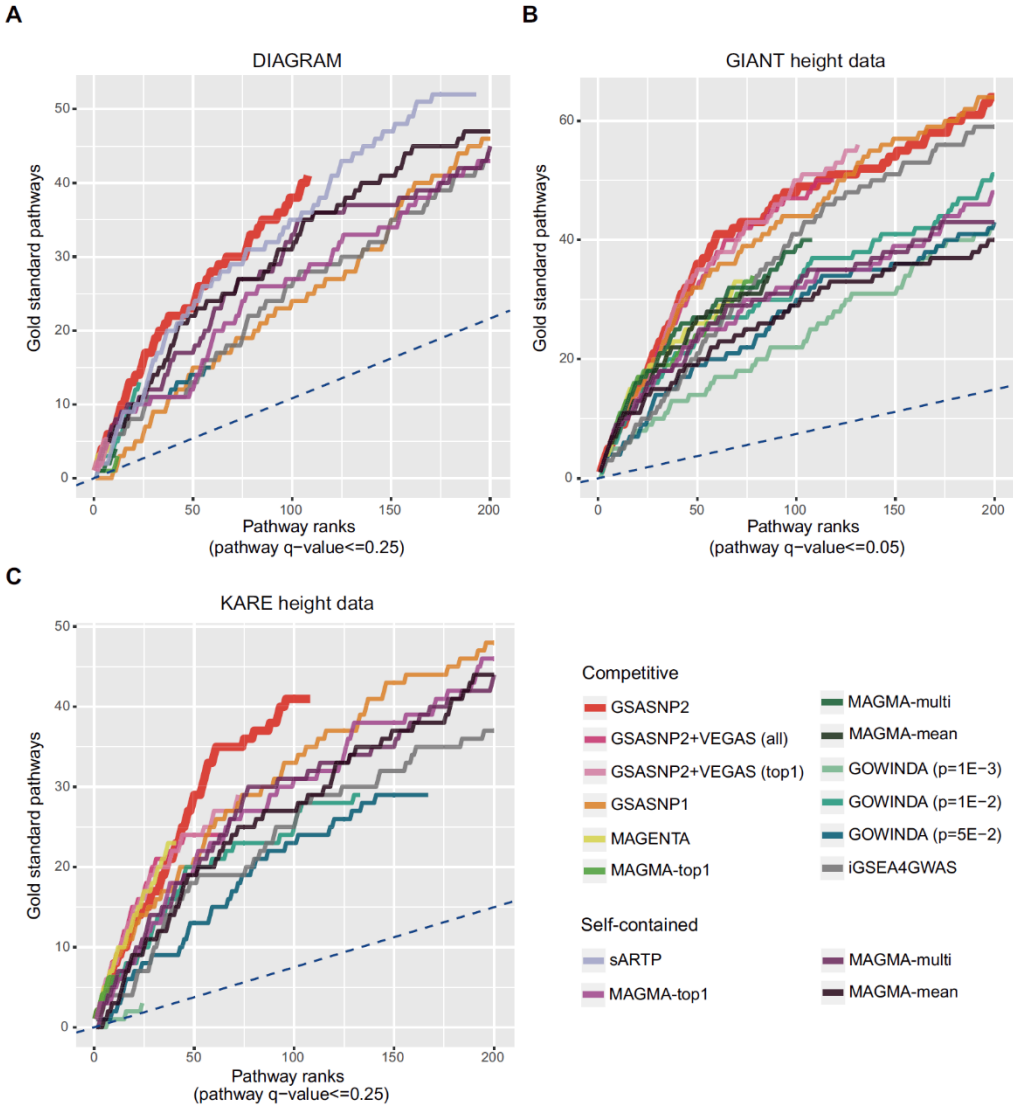
Set	Size	Count	z-score	Adj. z-sc	Adj. p-va	Adj. q-va	List of Genes
KEGG_MATURITY_ONSET_DIABETES_OF_THE_YOUNG	25	25	6.48959	6.83566	0	1E-08	HHEX (6.949345); HNF1A (4.566419); HI
REACTOME_REGULATION_OF_BETA_CELL_DEVELOPMENT	30	30	3.22243	4.71386	1.2E-06	0.00075	HNF1A (4.566419); HNF1B (4.105567); S
REACTOME_PRE_NOTCH_TRANSCRIPTION_AND_TRANSLATION	29	22	3.94822	4.71583	1.2E-06	0.00075	NOTCH4 (2.637982); E2F3 (2.469393); N
KEGG_ALLOGRAFT_REJECTION	38	35	5.64573	4.46046	4.1E-06	0.00127	HLA-DQA1 (4.694312); HLA-DRB1 (4.41
KEGG_GRAFT_VERSUS_HOST_DISEASE	42	38	4.90796	4.38941	5.7E-06	0.00142	HLA-DQA1 (4.694312); HLA-DRB1 (4.41
REACTOME_REGULATION_OF_GENE_EXPRESSION_IN_BETA_CELLS	20	20	2.30829	4.25435	1E-05	0.00218	HNF1A (4.566419); SLC2A2 (1.764304);
PID_HNF3B_PATHWAY	45	45	2.49585	4.24102	1.1E-05	0.00218	HNF1A (4.566419); HNF1B (4.105567); K
KEGG_TYPE_I_DIABETES_MELLITUS	44	42	5.75436	4.25228	1.1E-05	0.00218	HLA-DQA1 (4.694312); HLA-DRB1 (4.41
KEGG_TYPE_II_DIABETES_MELLITUS	47	45	4.11348	4.24052	1.1E-05	0.00218	KCNJ11 (4.023013); ABCC8 (3.506411);
BIOCARTA_CELLCYCLE_PATHWAY	23	23	2.89012	4.06483	2.4E-05	0.003	CDKN2B (3.982983); CDKN2A (3.53022(
BIOCARTA_VIP_PATHWAY	29	25	2.7754	4.00291	3.1E-05	0.00355	PLCG1 (3.209592); CHUK (3.166127); NF
PID_LYSOPHOSPHOLIPID_PATHWAY	66	65	3.61517	3.98202	3.4E-05	0.00355	ADCY5 (3.706311); BCAR1 (3.637863); P
KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_BIOSYNTHESIS	11	10	3.4107	3.83483	6.3E-05	0.00603	PDHA2 (3.211462); VARS (1.872344); LA
REACTOME_CYTOSOLIC_TRNA_AMINOACYLATION	24	24	2.53806	3.75272	8.7E-05	0.00779	RARS (3.917630); CARS (2.632525); EEF1
REACTOME_NOTCH_HLH_TRANSCRIPTION_PATHWAY	13	11	3.25256	3.7318	9.5E-05	0.0079	NOTCH4 (2.637982); NOTCH2 (2.03075
REACTOME_POTASSIUM_CHANNELS	98	97	4.96274	3.61476	0.00015	0.01171	KCNQ1 (4.738121); KCNJ11 (4.023013);
KEGG_NOTCH_SIGNALING_PATHWAY	47	47	2.11343	3.60703	0.00015	0.01171	NOTCH4 (2.637982); NOTCH2 (2.03075
PID_AP1_PATHWAY	70	68	10.4268	3.53142	0.00021	0.01368	TCF7L2 (6.252759); CDKN2A (3.530220);
PID_NOTCH_PATHWAY	59	59	2.62607	3.54338	0.0002	0.01368	NOTCH4 (2.637982); FBXW7 (2.588611);
KEGG_ASTHMA	30	28	4.71906	3.49211	0.00024	0.01494	HLA-DQA1 (4.694312); HLA-DRB1 (4.41
REACTOME_INTEGRATION_OF_ENERGY_METABOLISM	120	110	3.49163	3.42256	0.00031	0.01842	KCNJ11 (4.023013); ADCY5 (3.706311);
REACTOME_RIP_MEDIATED_NFKB_ACTIVATION_VIA_DAI	18	17	2.00516	3.27502	0.00053	0.02994	CHUK (3.166127); AGER (2.398901); NFI
BIOCARTA_AKAP95_PATHWAY	12	12	1.54971	3.2586	0.00056	0.02994	PRKACG (1.742986); PRKACB (1.675484
KEGG_VIRAL_MYOCARDITIS	73	68	4.9149	3.26646	0.00054	0.02994	HLA-DQA1 (4.694312); HLA-DRB1 (4.41
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	89	81	3.7768	3.23911	0.0006	0.02994	HLA-DQA1 (4.694312); HLA-DRB1 (4.41
REACTOME_PKA_MEDIATED_PHOSPHORYLATION_OF_CREB	18	16	3.15014	3.22467	0.00063	0.03024	ADCY5 (3.706311); ADCY7 (2.166519); P
REACTOME_TRAF6_MEDIATED_NFKB_ACTIVATION	21	20	2.12798	3.21201	0.00066	0.03044	CHUK (3.166127); AGER (2.398901); NFI
BIOCARTA_IL5_PATHWAY	10	10	3.38768	3.18371	0.00073	0.03238	HLA-DRB1 (4.415043); HLA-DRA (3.101
BIOCARTA_RNA_PATHWAY	10	10	1.92901	3.15412	0.0008	0.03461	CHUK (3.166127); NFKB1 (1.923206); TP
REACTOME_PRE_NOTCH_EXPRESSION_AND_PROCESSING	44	37	2.4246	3.1376	0.00085	0.0354	NOTCH4 (2.637982); E2F3 (2.469393); N
KEGG_AUTOIMMUNE_THYROID_DISEASE	53	50	5.03694	3.11283	0.00093	0.03727	HLA-DQA1 (4.694312); HLA-DRB1 (4.41
BIOCARTA_G1_PATHWAY	28	28	2.40209	3.10072	0.00097	0.03761	CDKN2B (3.982983); CDKN2A (3.53022(
ST_GAQ_PATHWAY	28	26	1.846	3.08999	0.001	0.03769	CFB (3.240930); NFKBIL1 (3.173078); PL
KEGG_DILATED_CARDIOMYOPATHY	92	87	4.04287	3.091	0.001	0.03769	ADCY5 (3.706311); TNF (3.325753); ITG
REACTOME_ION_TRANSPORT_BY_P_TYPE_ATPASES	34	30	3.27508	3.06071	0.0011	0.03934	ATP2C1 (1.789229); ATP9A (1.481776);
BIOCARTA_TID_PATHWAY	19	19	1.90726	3.04395	0.00117	0.04023	TNF (3.325753); HSPA1A (2.215279); NF
KEGG_SMALL_CELL_LUNG_CANCER	84	81	3.67783	3.04551	0.00116	0.04023	CDKN2B (3.982983); CHUK (3.166127);
KEGG_PROSTATE_CANCER	89	85	9.68336	3.02374	0.00125	0.04097	TCF7L2 (6.252759); CHUK (3.166127); C
REACTOME_REGULATION_OF_INSULIN_SECRETION	93	85	3.16827	3.00162	0.00134	0.04223	KCNJ11 (4.023013); ADCY5 (3.706311);

Tests for real GWAS data

70K

180K

9K

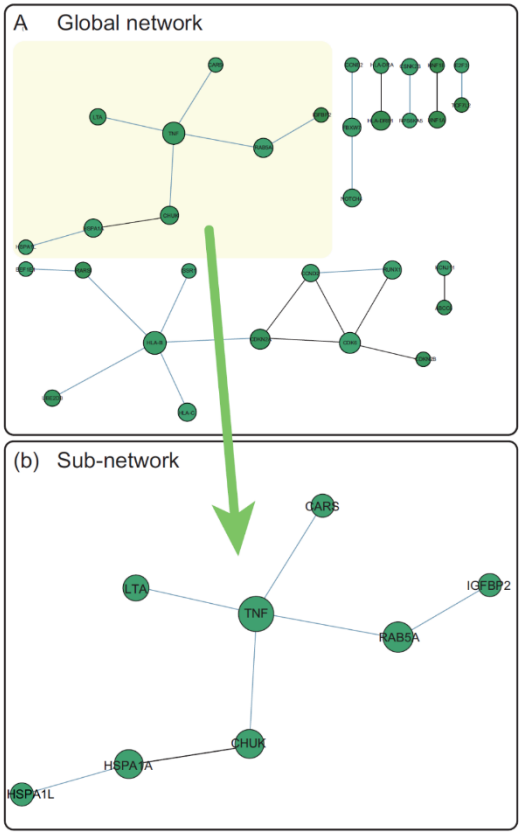


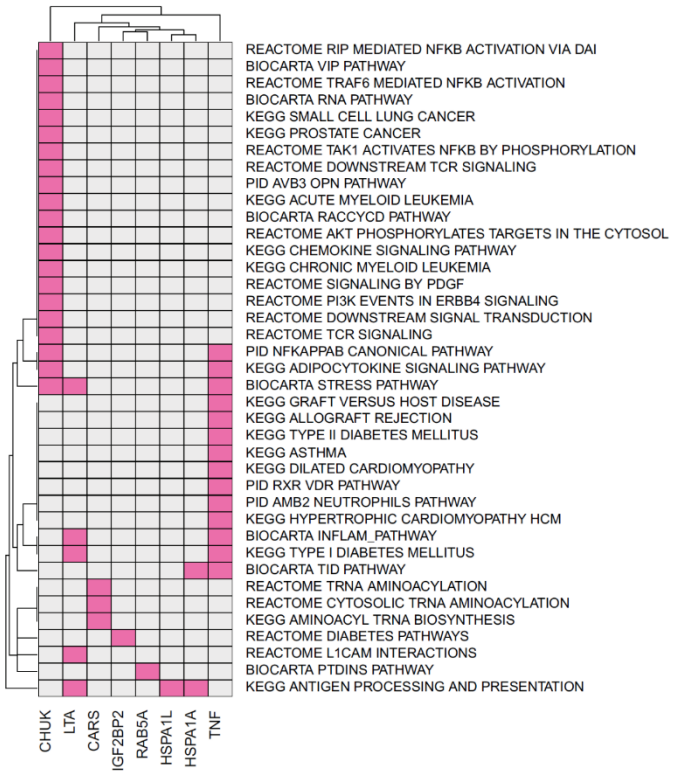
Computing time

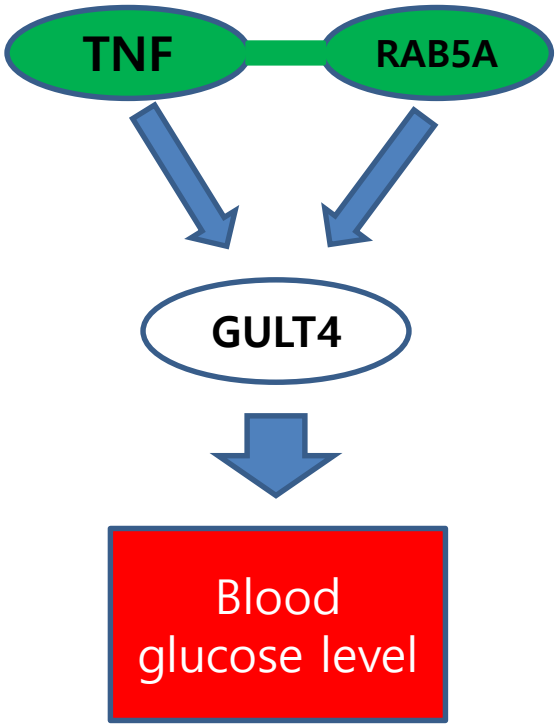
Table 2 Running times for eight pathway analysis programs for GWAS summary data

Method	Time	Permutation
GSA-SNP2	1.53 min	
GSA-SNP1	1.49 min	
MAGMA-mean	3.03 min	
MAGMA-top1	34.85 min	
MAGMA-multi	41.85 min	
iGSEA4GWAS	30 min	
MAGENTA	114.18 min	10 000
Gowinda ($P = 0.001$)	0.62 min	10 000
Gowinda ($P = 0.01$)	0.80 min	10 000
Gowinda ($P = 0.05$)	2.01 min	10 000
INRICH ($P1 = 1E\text{-}6$)	0.85 min	10 000
INRICH ($P1 = 1E\text{-}4$)	2.41 min	10 000
sARTP	10.41 days	100 000

Network analysis







Increasing power of GWAS: Meta-analysis

- Combining *p*-values from independent experiments
  - **Fisher's method**: $X^2_{2k} \sim -2\sum_{i=1}^{k} \ln(p_i)$
  - **Stouffer's method**: Z-scores rather than p-values, allowing incorporation of study weights

$$Z \sim \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$$

Increasing power of GWAS: Meta-analysis

	Fixed effect	**Random effect**
Inputs	β_i- effect size estimate for study i SE_i- standard error for study i	
Intermediate Statistics	$w_i = \frac{1}{SE_i{}^2}$ $SE = \sqrt{\frac{1}{\sum_i w_i}}$ $\beta = \sum_i \frac{\beta_i w_i}{\sum_i w_i}$	$\tau^2 = \frac{Q - N_i + 1}{c}$ $w_i^* = \frac{1}{SE_i{}^2 + \tau^2}$ $SE = \sqrt{\frac{1}{\sum_i w_i^*}}$, $\beta = \sum_i \frac{\beta_i w_i^*}{\sum_i w_i^*}$
Meta Z-score	$Z = \frac{\beta}{SE}$	
Meta p-value	$P = 2\Phi(\lvert -Z \rvert)$	

More powerful **More conservative**

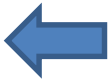
Pitfalls of existing methods

- Combines all the given *p*-values (or effect sizes). This is not always beneficial because…
  - Some cohorts may not be associated
  - Some cohort data may have low qualities
- Can we select only 'associated' cohorts for each SNP and integrate their *p*-values?

ORDMETA method

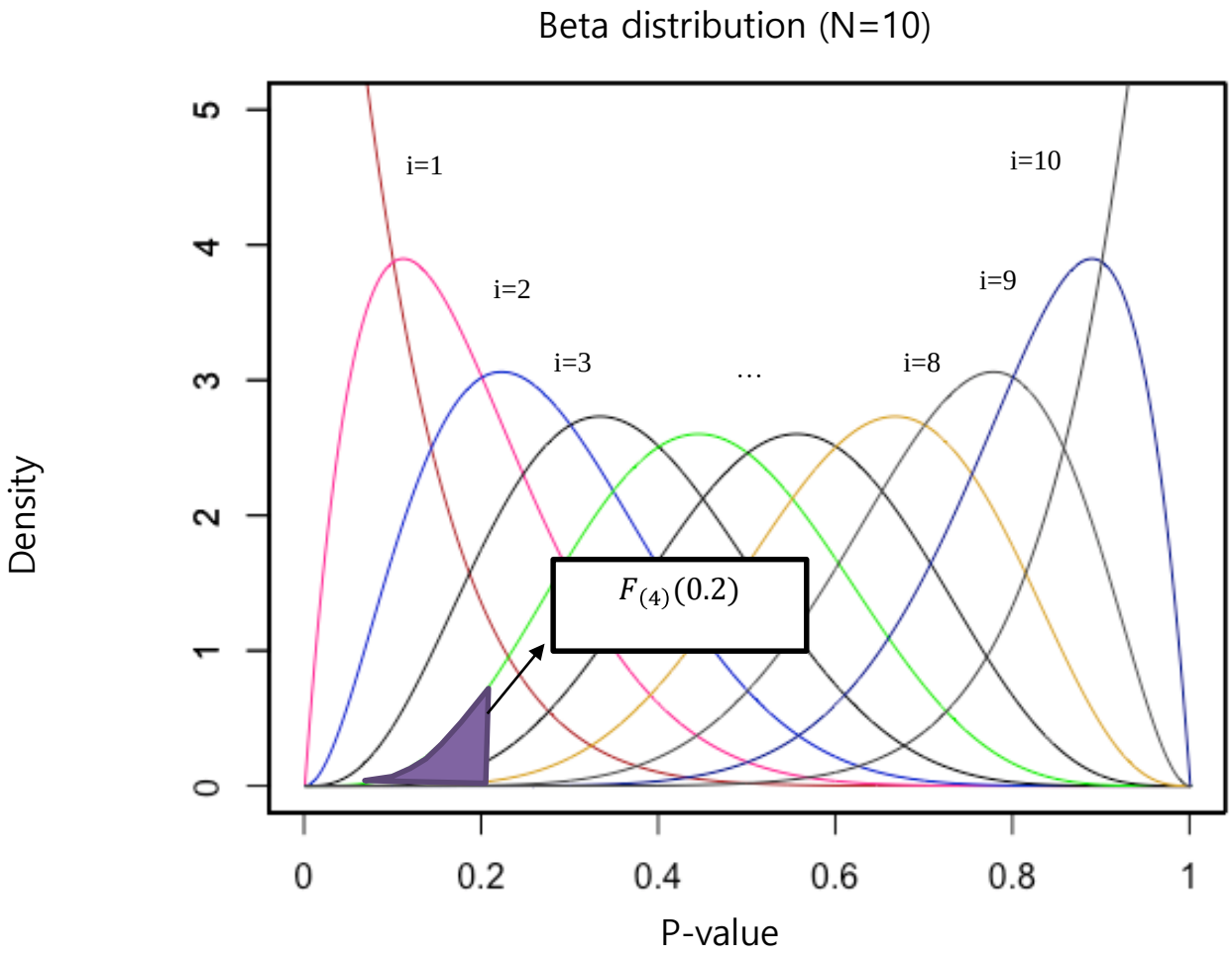
- Combines *p*-values from independent experiments based on joint order distribution
- $p_i \sim Unif(0,1)$ (*i*=1,…,N) : each independent *p*-value has a uniform distribution
- $(p_{(1)}, p_{(2)}, \ldots, p_{(N)})$: *N* ordered *p*-values have joint order distribution where each ordered *p*-value has a beta distribution

$$p_{(i)} \sim Beta(i, N - i + 1)$$



Marginal distribution

ORDMETA method



ORDMETA method

- We calculate '***p*-value for the minimum marginal *p*-value**' of the joint order distribution

Example: GIANT-BMI data analysis

1	SNPNAME	P_African	P_EastAsian	P_Hispanic	P_SouthAsian	P_ORDMETA	P_METAL	P_Fixed	P_Random	ORDMETA_C
8	rs7561317	0.0067	0.067	0.0023	0.014	9.97E-06	8.88E-07	1.21E-06	1.21E-06	1 2 3 4
9	rs7647305	4.30E-05	0.14	0.023	0.033	0.0001387	2.90E-07	2.19E-07	2.19E-07	1 3 4
10	rs10938397	2.40E-06	0.22	0.042	0.093	3.78E-05	4.40E-07	1.83E-06	0.002376	1
11	rs2206277	2.30E-06	0.35	6.00E-04	0.0067	1.19E-06	4.13E-08	4.79E-08	0.022122	1 3 4
12	rs987237	0.00067	0.11	0.00071	0.0037	2.02E-07	4.96E-06	3.49E-06	0.073536	1 3 4
13	rs2301680	0.05	0.018	0.092	9.50E-05	3.53E-05	4.92E-07	3.52E-07	3.52E-07	1 2 3 4
14	rs7901695	0.048	0.95	0.0022	1.40E-07	2.22E-06	2.47E-08	4.85E-09	0.00156	4
15	rs4506565	0.056	0.81	0.00021	6.30E-08	5.27E-07	6.38E-09	1.04E-09	0.003102	3 4
16	rs7903146	0.00052	0.82	3.90E-06	5.60E-08	1.82E-10	1.24E-12	9.21E-14	6.48E-06	3 4
17	rs12243326	0.58	0.78	5.60E-06	1.30E-05	2.03E-09	3.10E-05	4.01E-05	0.094958	3 4
18	rs12429545	0.00019	0.27	0.097	0.0046	0.00024626	5.96E-07	9.58E-06	0.000809	1 4
19	rs9930333	0.16	0.00072	1.90E-06	1.50E-09	4.33E-11	5.38E-14	2.76E-13	0.000979	3 4
20	rs1421085	1.40E-06	0.004	1.90E-08	1.50E-12	0	3.63E-25	6.76E-26	7.19E-24	1 3 4
21	rs1558902	1.70E-06	0.004	2.40E-08	5.90E-13	0	2.36E-25	3.92E-26	5.08E-26	1 3 4

Green: significant for the p-value 10^{-6}

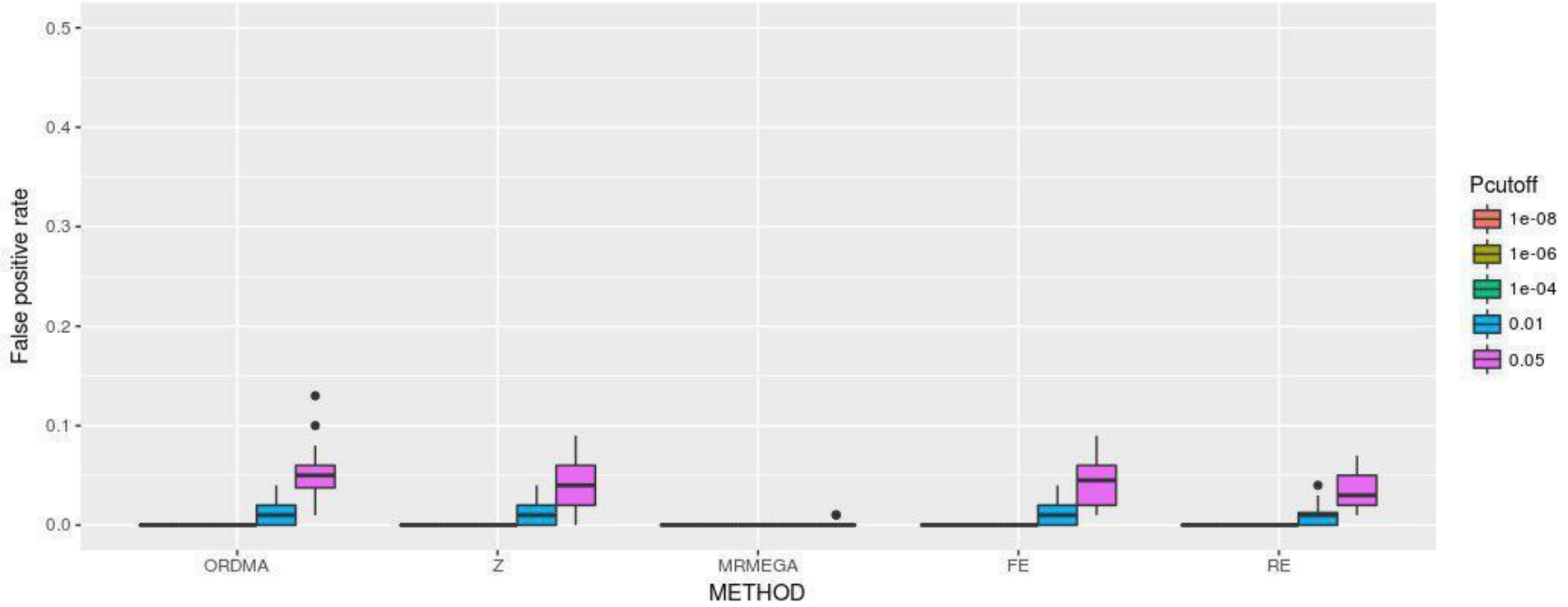
Simulation for false positive control test

- **15 cohorts**, each containing 2000 samples generated from 1000 Genomes data

- **100 Effect SNPs**
  - Effect SNP 1~40: common to all 15 cohorts
  - Effect SNP 41~60: specific to cohort 1~5
  - Effect SNP 61~80: specific to cohort 6~10
  - Effect SNP 81~100: specific to cohort 11~15

- **Very small heritability=1E-5**
- Compared methods: ORDMETA, METAL, fixed effect, random effect and MR-MEGA

Simulation test: comparison of false positive controls



Simulation for power test

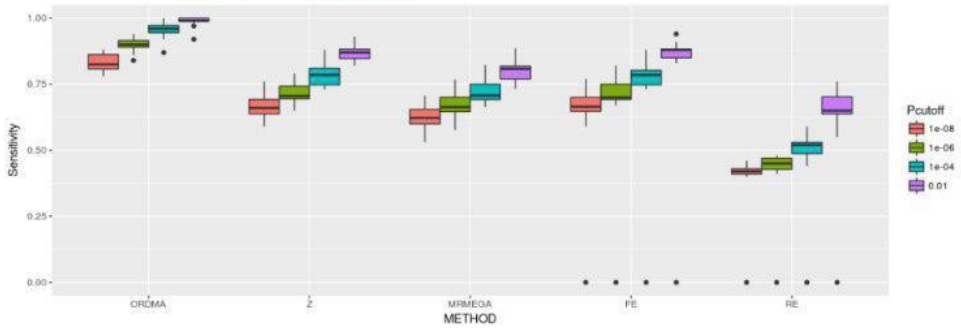
- **15 cohorts**, each contained 2000 samples generated from 1000 Genomes data

- **100 Effect SNPs**
  - Effect SNP 1~40: common to all 15 cohorts
  - Effect SNP 41~60: specific to cohort 1~5
  - Effect SNP 61~80: specific to cohort 6~10
  - Effect SNP 81~100: specific to cohort 11~15

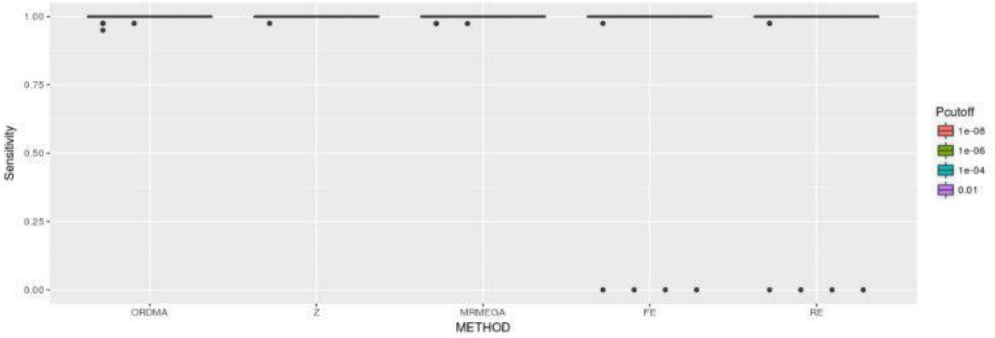
- **Heritability=0.5**
- **So, 60% of the effect SNPs are 'associated' in only five cohorts**

Simulation test: comparison of powers

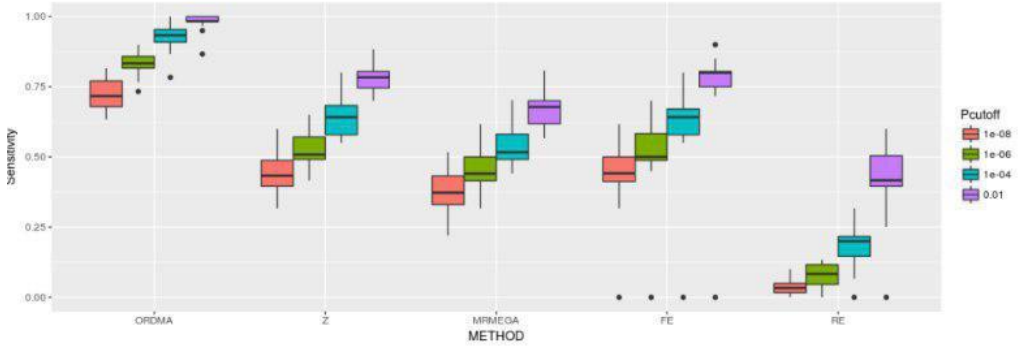
All Effect SNPs



Common Effect SNPs

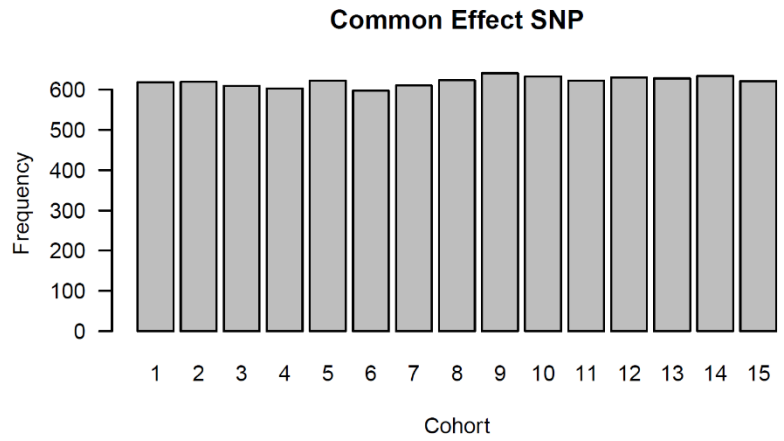


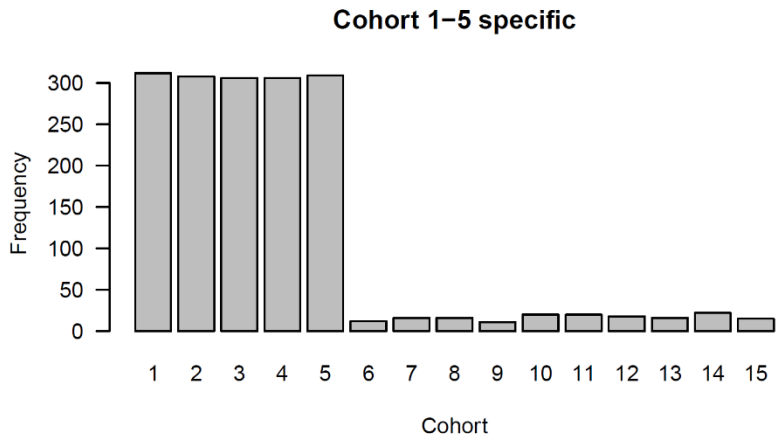
Group-specific Effect SNPs

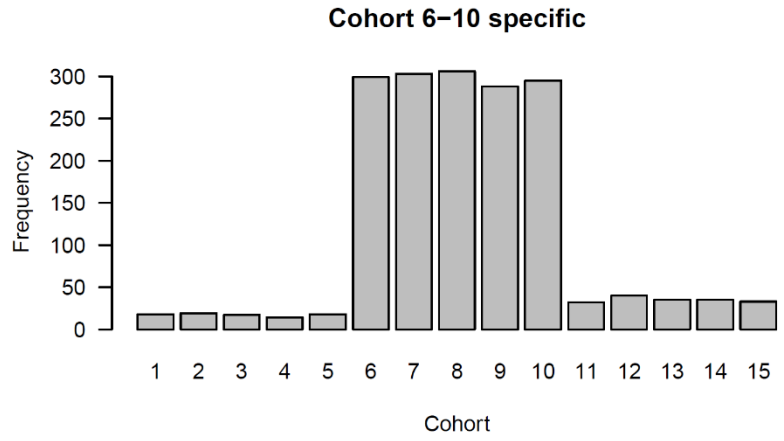


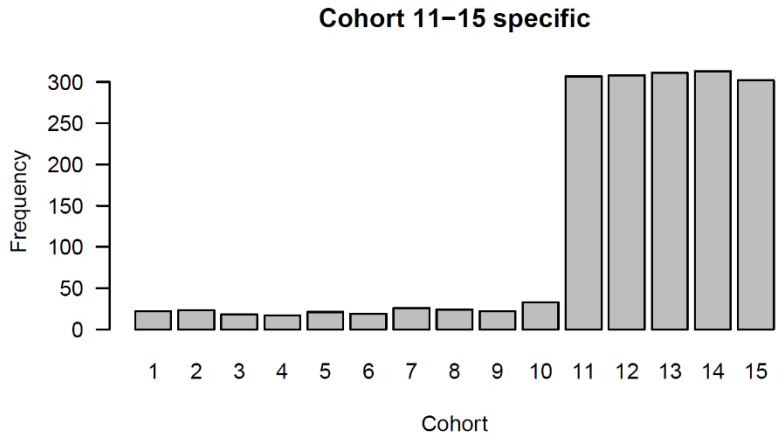
- 15 cohorts (each contains 2000 samples).
- 100 Effect SNPs
  - SNP 1~40: common to all cohorts
  - SNP 41~60: Specific to cohort 1~5
  - SNP 61~80: Specific to cohort 6~10
  - SNP 81~100: Specific to cohort 11~15
- Heritability=0.5
- 20 repetition

Simulation test: detection of associated cohorts









Many Thanks!

- Our Lab
  - Hai, C. T. Nguyen
  - Sora Yoon
  - Jinhwan Kim
  - Juok Cho
  - Soungou Kim
  - Bukyung Baik
- SNU
  - Prof. Yun Joo Yoo