A meta-analysis of genome-wide association with body mass index

Name: Guojun Ma

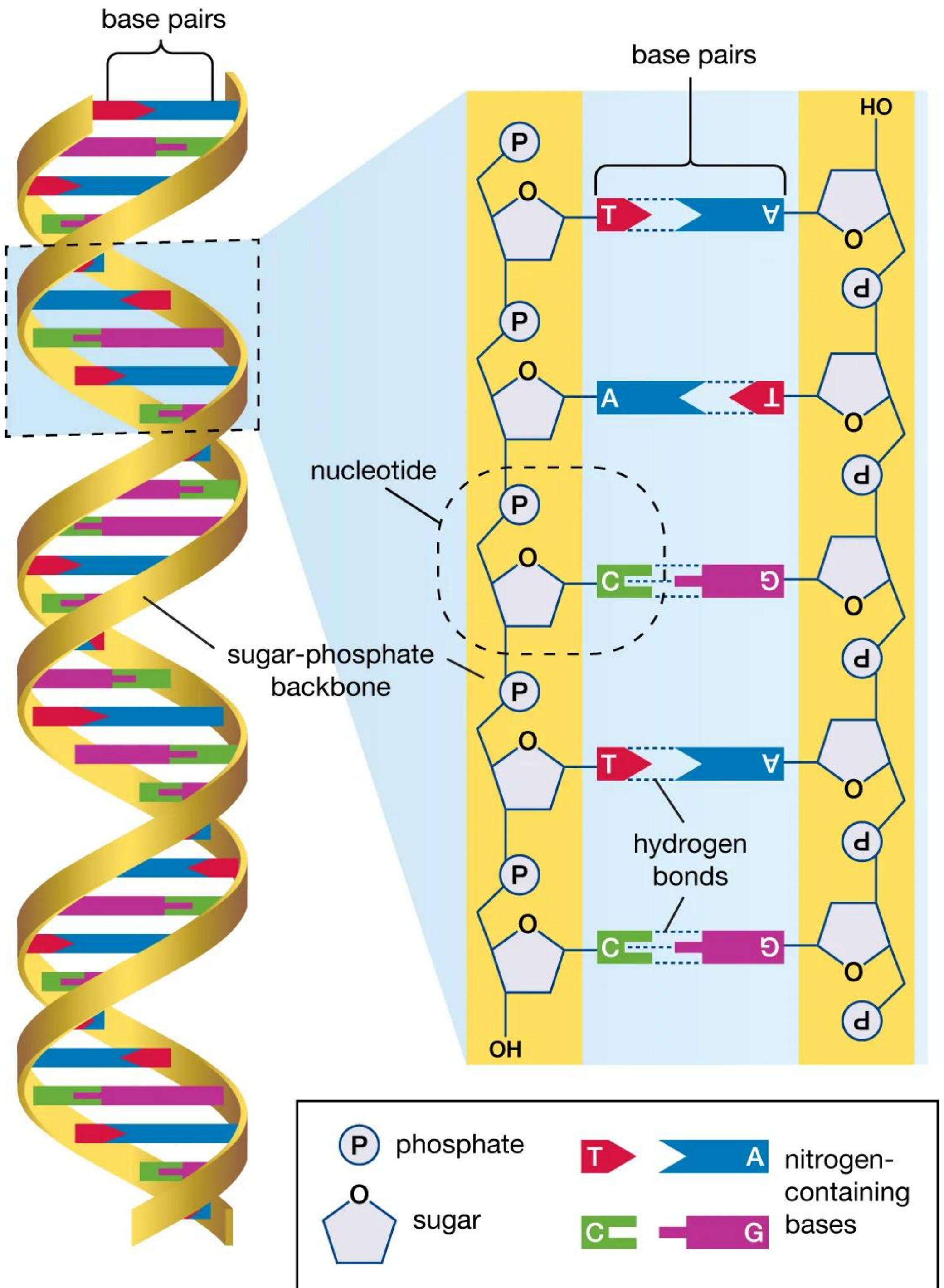
Supervisor: Dr. Yu-ting Chen and Dr. Xuekui Zhang

Introduction

- In this project, we aim to identify the genetic variations in the human genome underlying body mass index(BMI), by conducting a meta-analysis of previously published studies.
- While previous meta-analysis focused solely on the European population, our study encompasses larger and more diverse samples.
- The objectives are discovering previously unknown genetic variations associated with BMI, as well as comparing the impact of these genetic variations across distinct ethnicity group.
- Our research adds to the expanding knowledge regarding the genetic underpinnings of BMI, with the potential for valuable medical applications.

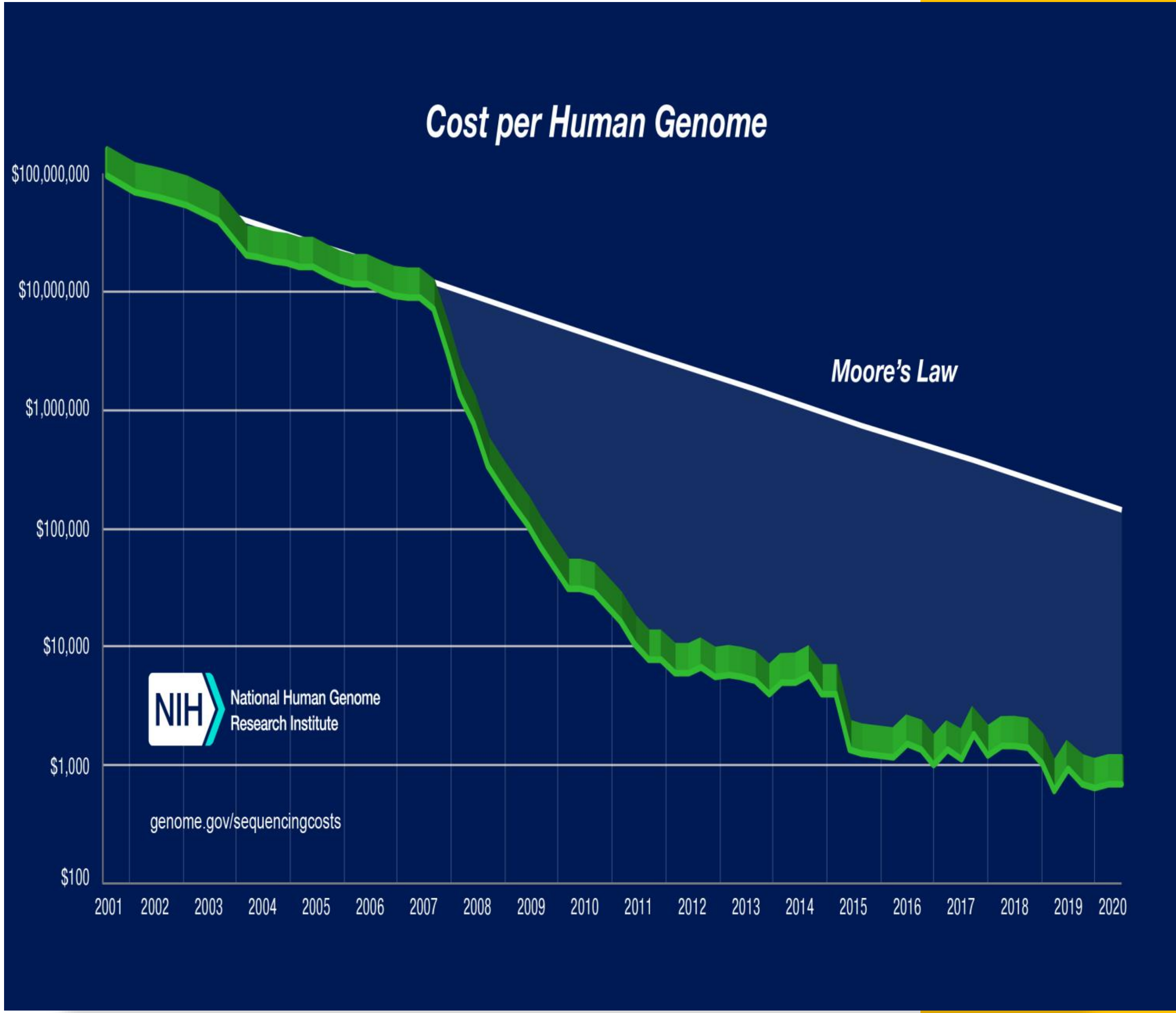
Background

- The human genome comprises around three billion base pairs of deoxyribonucleic acid (DNA), held together by two strands of sugar-phosphate backbones.
- Each DNA has four different nucleotide types: Adenine(A), Thymine(T), Guanine(G) and Cytosine(C).
- Numerous combination of DNA encode distinct information, and variations in these combinations lead to diverse phenotypes.
- The **single nucleotide polymorphisms (SNPs)** are the most common form of genetic variations throughout populations(> 10% in the population). The possible nucleotide variations of the SNP are referred to as **allele**.



GWAS study

- Genome-wide association study (GWAS) aims to find the relationship between genetic variations (SNPs) and susceptibility to diseases or traits across the entire genome.
- The first GWAS study was conducted by the Wellcome Trust case control Consortium(WTCCC) in 2005, shortly after the completion of Human Genome Project.
- In the last decade, GWAS has gained substantial popularity, primarily attributed to the declining costs of genome sequencing.
- By the year 2017, Researchers have identified approximately 55,000 unique loci in the genome associated with many traits and diseases.



GWAS study

- GWAS generally require a large sample size. Collected data from study cohorts or use available genetic and phenotypic information from biobanks or repositories.
- Genotypic data can be collected using microarrays to capture common variants, or next-generation sequencing methods for whole-genome sequencing(WGS) or whole-exome sequencing sequencing(WES) that also include rare variants.
- Reliable results from GWAS requires careful quality control, such as removing rare variants, removing variants that are not in Hardy-Weinberg equilibrium, identifying and removing genotyping errors and among others.
- Untyped genotypes imputed using information from matched reference population from repositories such as 1000 Genomes Project.

Association testing

- Genetic association tests are performed for each SNP using an appropriate model, such as the mixed linear regression model.

$$y_i = \beta G_i + \gamma A_i + e_i,\ e_i \sim N(0, \sigma^2),$$

- Where y_i is the quantitative phenotype for the i-th individual; $G_i = \{0,1,2\}\ depends\ on\ the\ phenotype\ of\ the\ ith\ invidual;\ A_i$ encode the characteristic of the i-th individual such as age, sex, location and principal components of genotype; β is the genetic effect of the SNP.
- The p-value is obtained by testing the null hypothesis $\beta = 0 \text{ vs the alternative } \beta \neq 0.$
- Due to the extensive number of SNPs involved (approximately 1 million), it is crucial to control the false positive rate. The Bonferroni correction is frequently employed, wherein the p-value threshold is set at 5×10^{-8}.

Meta-analysis

- Meta-analysis is commonly employed to integrate the findings of multiple GWAS studies through a statistical framework.
  - It increases the statistical power of GWAS study with a larger sample size;
  - Enhances the precision and robustness of research findings;
  - Examines the cross-ethnicity replicability and variability of genetic effects.
- There are multiple types of Meta-analysis:
  - Patient-level: collect and analysis the genotyped data from multiple cohorts.
  - Summary-level: collect only the summary statistics from previous studies.

Methodology

- We conducted a summary-level meta-analysis on the recently published studies on BMI.
- We conducted a thorough literature search on the GWAS catalog website, a comprehensive database that compiles data of published GWAS.
- We identified a few studies published in recent years (2015 – 2022) for which the summary statistics are publicly accessible. The summary statistics include the association level of millions of SNPs.

Cohorts

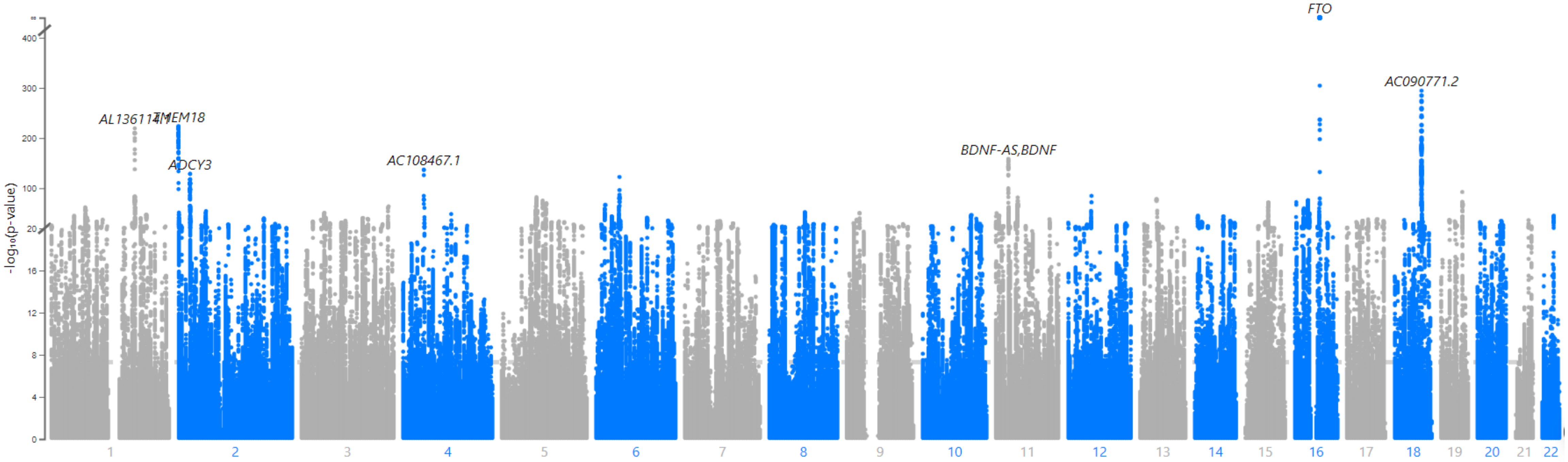
Cohorts	Sample sizes	Ethnicity	Year of publication	Reference
UK Biobank + GIANT	694,649	White European	2016	Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry
Korea biobank	72,298	East Asian	2022	Genome-wide study on 72,298 individuals in Korean biobank data for 76 traits
Taiwan Biobank	21,930	East Asian	2022	Genome-wide association study identifies genetic risk loci for adiposity in a Taiwanese population.
Japan biobank	179,000	East Asian	2021	A cross-population atlas of genetic associations for 220 human phenotypes
Hispanic/Latino Anthropometry (HISLA) Consortium	56,161	Hispanic/Latino	2022	Ancestral diversity improves discovery and fine-mapping of genetic loci for anthropometric traits-The Hispanic/Latino Anthropometry Consortium

Methodology

- We performed the fixed-effect inverse-variance weighted meta-analysis, using the command line software METAL.
  - Fixed-effect: assume the genetic effects are consistent across all cohorts.
  - Inverse-variance: assigning the weight to the effect size from each study based on the inverse of its variance.
- $\beta_i: effect\ size\ estimate\ from\ study$ i;
  $se_i: standard\ error\ of\ estimate\ from\ study\ i$;
- The overall effect estimate is $\beta = \frac{\sum_i \beta_i w_i}{\sum_i w_i}$, where $w_i = \frac{1}{se_i^2}$
- The overall standard error is $se = \sqrt{\frac{1}{\sum_i w_i}}$

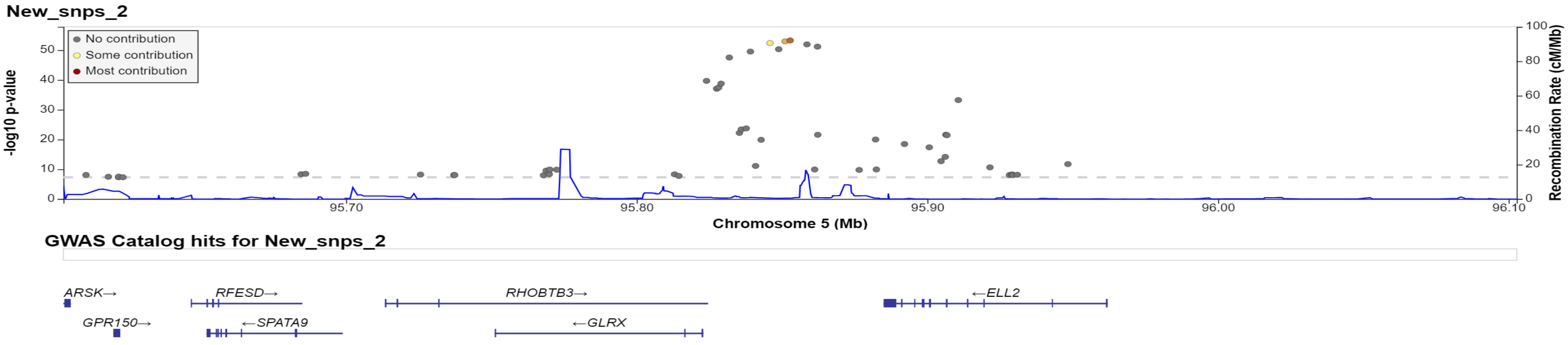
Results

- We analyzed a subset (~2.3 millions) of SNPs, which were reported in the previous meta-analysis study (UKB + GIANT cohorts).
- Using the strict threshold of the p-value (5×10^{-8}), we identified a total of 61,507 significant SNPs.
- If we break the genome into different parts (a disjoint window of 500Kb), these significant SNPs correspond to 1,966 different loci. For comparison, the previous meta-analysis identified 981 loci.

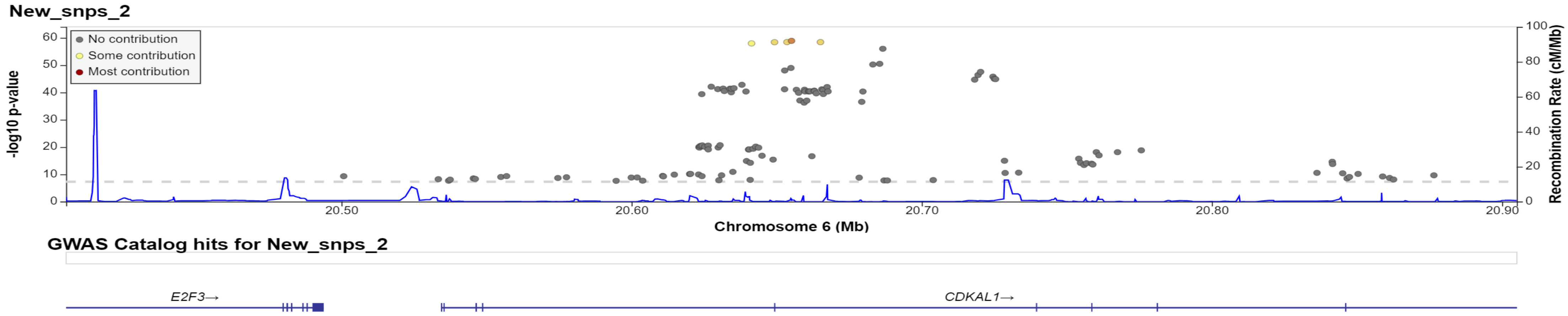


Regional plot

ELL2

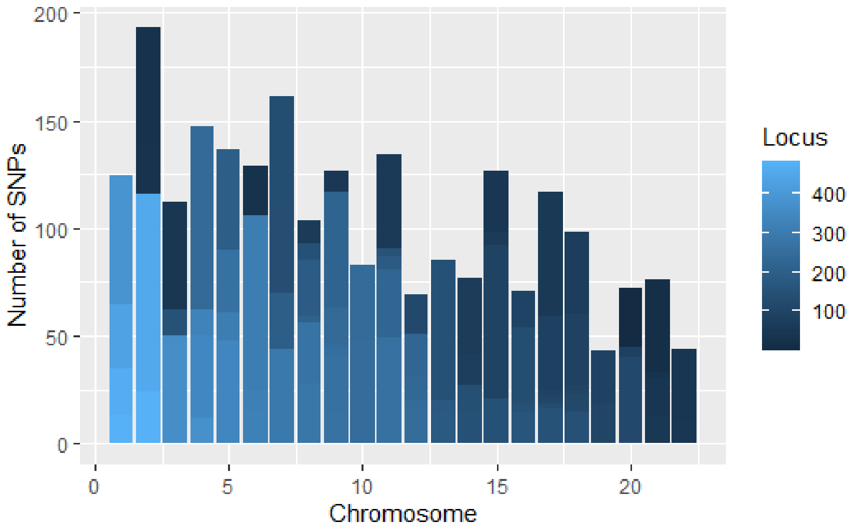


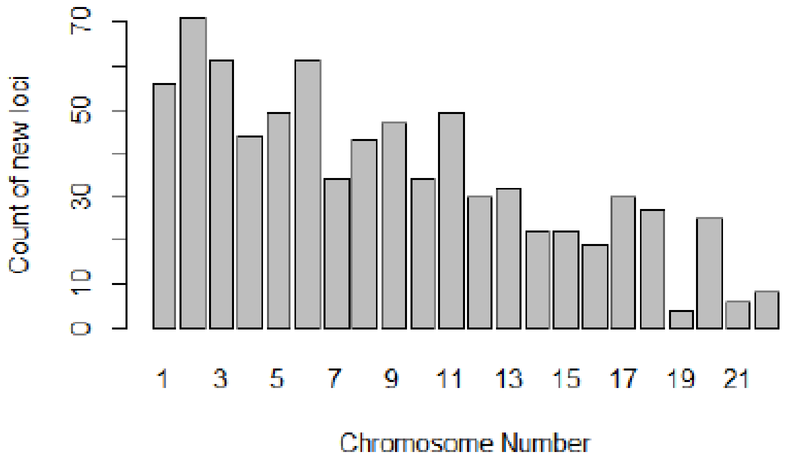
CDKAL1



Distribution of new significant SNPs

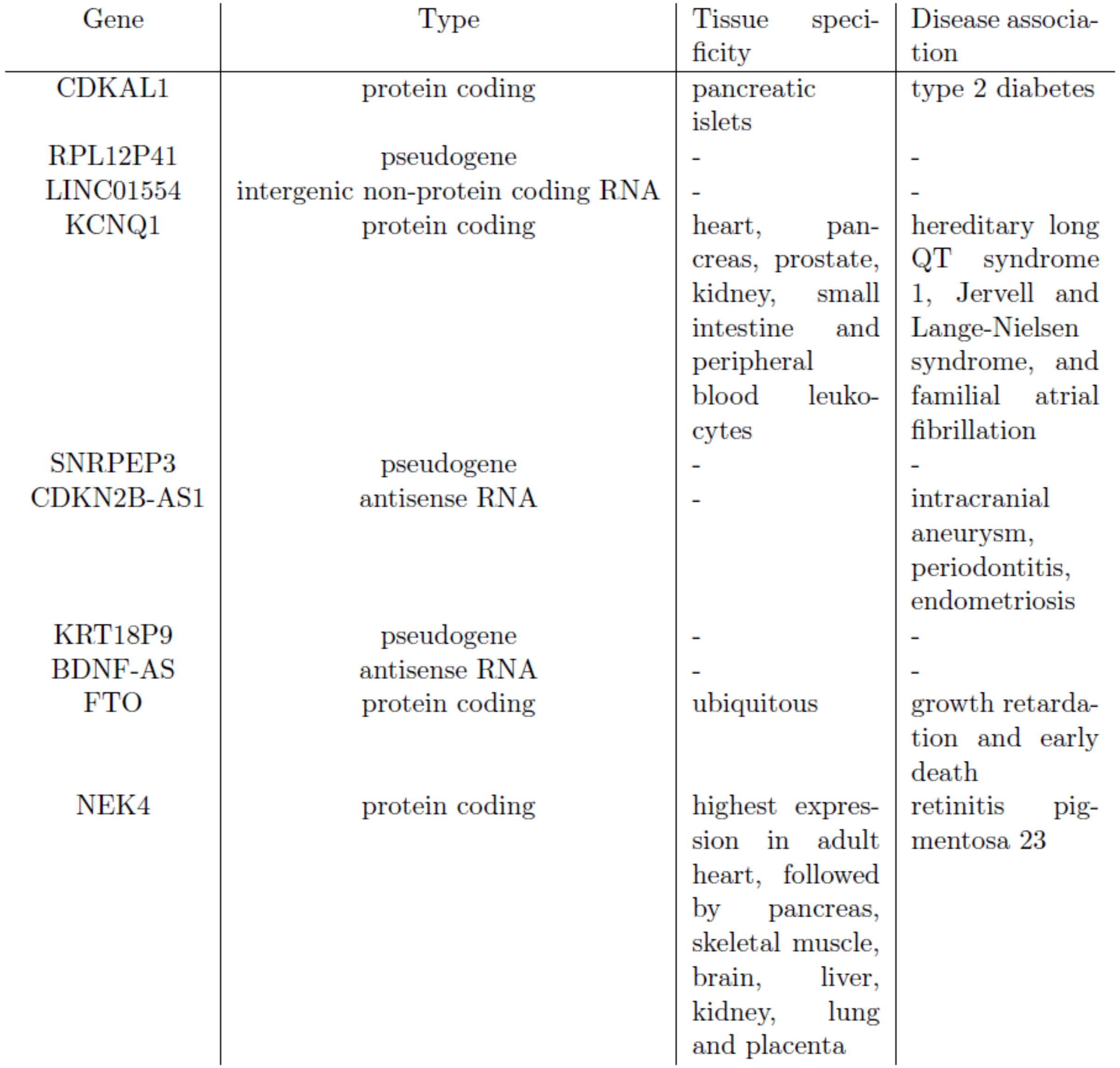
- In comparison to the previous GWAS study on the UKB+GIANT cohort, our meta-analysis has identified an additional 27,616 significant SNPs.
- These SNPs correspond to 774 new loci, each of which encompasses more than 10 significant SNPs.
- The plots show the distribution of those SNPs and loci.





Gene Functions

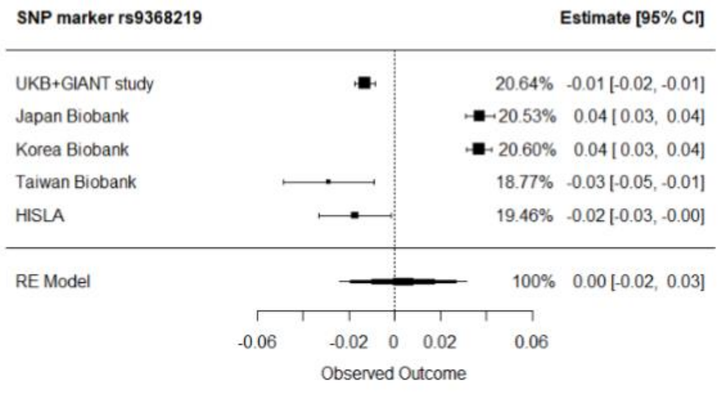
- We identified the nearby genes for the 10 new SNPs with the highest levels of significant association.
- The functions of these genes can be found using the UCSC Genome Browser gateway.
- We identified that several genes are the pseudogene which are no longer functional.



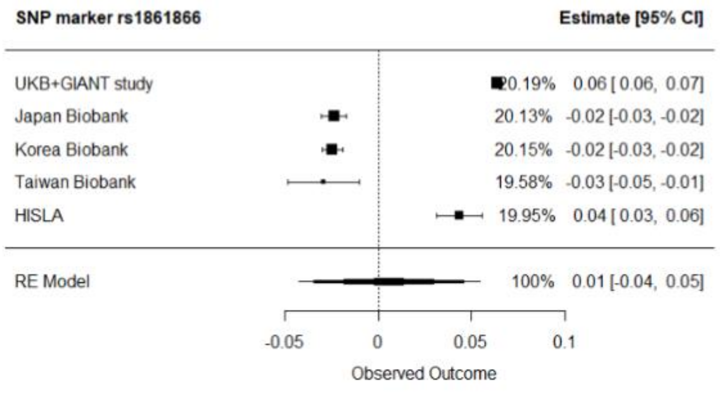
Gene	Type	Tissue specificity	Disease association
CDKAL1	protein coding	pancreatic islets	type 2 diabetes
RPL12P41	pseudogene	-	-
LINC01554	intergenic non-protein coding RNA	-	-
KCNQ1	protein coding	heart, pancreas, prostate, kidney, small intestine and peripheral blood leukocytes	hereditary long QT syndrome 1, Jervell and Lange-Nielsen syndrome, and familial atrial fibrillation
SNRPEP3	pseudogene	-	-
CDKN2B-AS1	antisense RNA	-	intracranial aneurysm, periodontitis, endometriosis
KRT18P9	pseudogene	-	-
BDNF-AS	antisense RNA	-	-
FTO	protein coding	ubiquitous	growth retardation and early death
NEK4	protein coding	highest expression in adult heart, followed by pancreas, skeletal muscle, brain, liver, kidney, lung and placenta	retinitis pigmentosa 23

Heterogeneity analysis

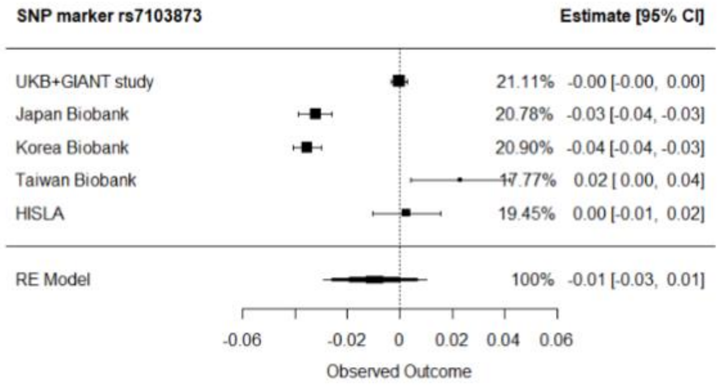
- Compare the effect size of SNP markers across different cohorts.
- Some SNPs exhibit heterogeneity between ethnicity groups. (~ 10 %)
- For example, SNPs rs9368219 and rs1861866 have opposite effect size.



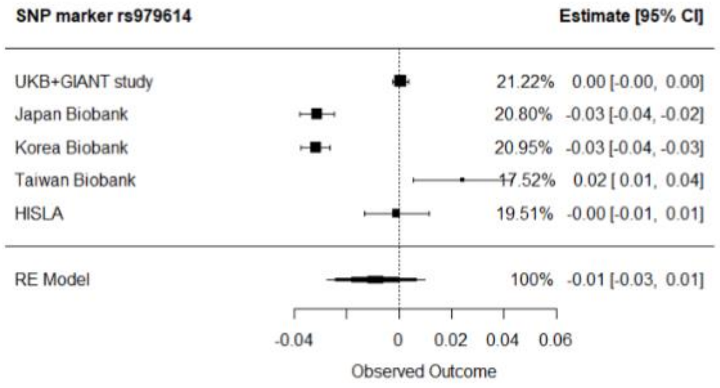
SNP marker rs9368219		Estimate [95% CI]
UKB+GIANT study	20.64%	-0.01 [-0.02, -0.01]
Japan Biobank	20.53%	0.04 [0.03, 0.04]
Korea Biobank	20.60%	0.04 [0.03, 0.04]
Taiwan Biobank	18.77%	-0.03 [-0.05, -0.01]
HISLA	19.46%	-0.02 [-0.03, -0.00]
RE Model	100%	0.00 [-0.02, 0.03]



SNP marker rs1861866		Estimate [95% CI]
UKB+GIANT study	20.19%	0.06 [0.06, 0.07]
Japan Biobank	20.13%	-0.02 [-0.03, -0.02]
Korea Biobank	20.15%	-0.02 [-0.03, -0.02]
Taiwan Biobank	19.58%	-0.03 [-0.05, -0.01]
HISLA	19.95%	0.04 [0.03, 0.06]
RE Model	100%	0.01 [-0.04, 0.05]



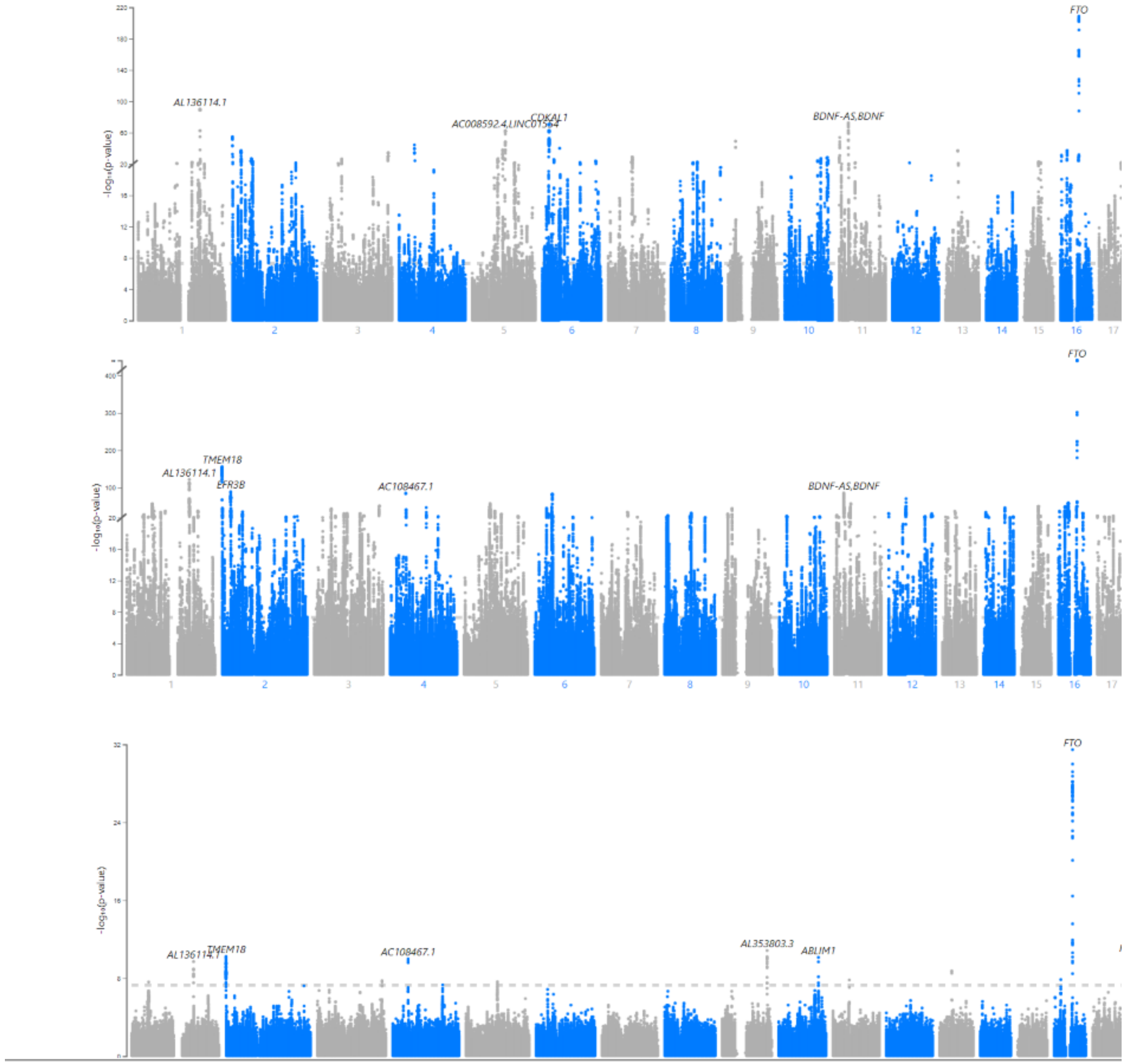
SNP marker rs7103873		Estimate [95% CI]
UKB+GIANT study	21.11%	-0.00 [-0.00, 0.00]
Japan Biobank	20.78%	-0.03 [-0.04, -0.03]
Korea Biobank	20.90%	-0.04 [-0.04, -0.03]
Taiwan Biobank	17.77%	0.02 [0.00, 0.04]
HISLA	19.45%	0.00 [-0.01, 0.02]
RE Model	100%	-0.01 [-0.03, 0.01]

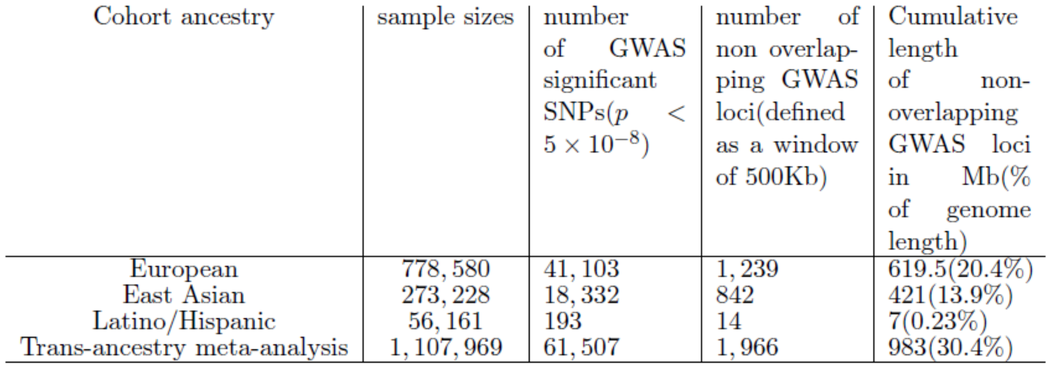


SNP marker rs979614		Estimate [95% CI]
UKB+GIANT study	21.22%	0.00 [-0.00, 0.00]
Japan Biobank	20.80%	-0.03 [-0.04, -0.02]
Korea Biobank	20.95%	-0.03 [-0.04, -0.03]
Taiwan Biobank	17.52%	0.02 [0.01, 0.04]
HISLA	19.51%	-0.00 [-0.01, 0.01]
RE Model	100%	-0.01 [-0.03, 0.01]

Subgroup analysis

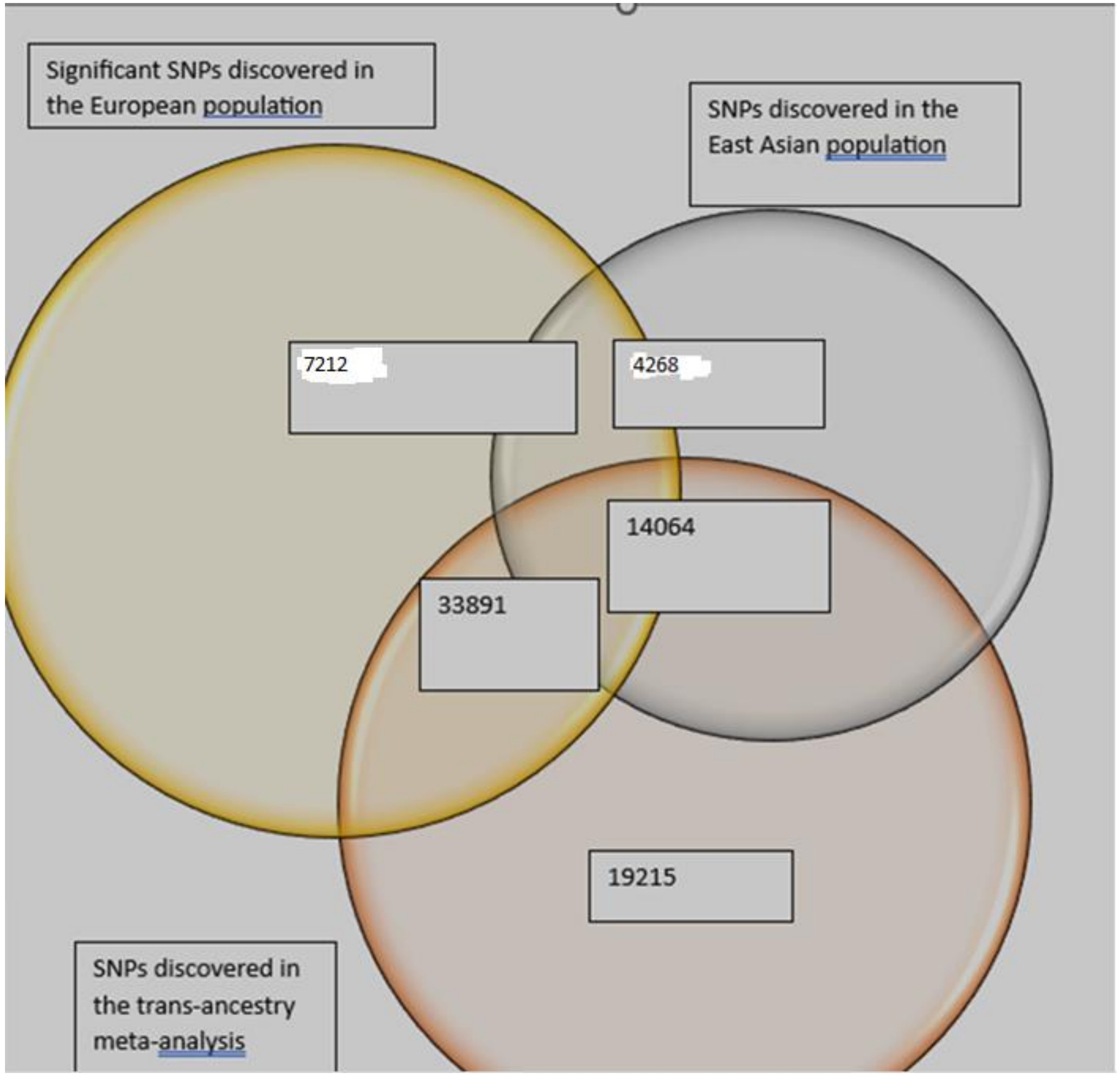
- Perform meta-analysis for each ethnicity group – to identify which SNPs are common in all population, as well as SNPs that are unique in the population.
- We identified that genes FTO, RSL24D1P11, AL136114.1 and many others (the loci with the highest association level) are common in each ancestry group.
- But also identified 8908 new SNPs(27 new loci) that are unique in the east Asian group, as well as 25 new SNPs(2 new loci) that are unique in the Latino/Hispanic population.





Cohort ancestry	sample sizes	number of GWAS significant SNPs($p < 5 \times 10^{-8}$)	number of non overlapping GWAS loci(defined as a window of 500Kb)	Cumulative length of non-overlapping GWAS loci in Mb(% of genome length)
European	778, 580	41, 103	1, 239	619.5(20.4%)
East Asian	273, 228	18, 332	842	421(13.9%)
Latino/Hispanic	56, 161	193	14	7(0.23%)
Trans-ancestry meta-analysis	1, 107, 969	61, 507	1, 966	983(30.4%)

The trans-ancestry meta-analysis confirmed the presence of many previously discovered SNPs, while also unveiling an additional 19,215 new SNPs.



Significant SNPs discovered in the European population
SNPs discovered in the East Asian population
7212
4268
14064
33891
19215
SNPs discovered in the trans-ancestry meta-analysis

Discussion

- Our research has unveiled a notable rise in the count of associated SNPs. However, it's important to acknowledge that these associations might be spurious due to confounding biases.
  - The genetic compositions of various ethnic groups differ significantly, a phenomenon referred to as population stratification. This can lead to certain SNPs displaying apparent associations even when there isn't a causal relationship.
  - SNPs in close proxy may be in linkage disequilibrium.
- Our findings indicate a substantial quantity of distinctive SNPs specific to the East Asian population, while there are considerably fewer unique SNPs in the Latino/Hispanic population.
  - Underpower of GWAS due to a small sample size.
- The polygenic nature of BMI(many signal with small effect size) presents a challenge for understanding the biological mechanisms and exploring potential therapeutic interventions.
  - Individuals with the same disease may have unique genetic profiles.

Future studies

- Incorporating a more diverse population into the meta-analysis to obtain a more comprehensive grasp of the genetic architecture.
- To achieve more robust results, it's essential to control for confounding biases like population stratification.
  - ➢ Existing methods possess certain limitations; for instance, LD score regression might exhibit inaccuracies when dealing with large sample sizes.
  - ➢ Developing a more effective correction for potential bias is a continuous focus within the ongoing research in this field.
- In order to gain mechanistic and biological insight, there is a need for novel techniques that deal with polygenicity and translate the finding of GWAS discoveries.
  - Identifying how the causal SNPs influence genes and establishing connections with physiological and cellular functions

Reference

- Klein et al., 2005] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., et al. (2005). Complement factor h polymorphism in age-related macular degeneration. Science, 308(5720):385–389.
- MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). Nucleic acids research, 45(D1):D896–D901.
- Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., Frayling, T. M., Hirschhorn, J., Yang, J., Visscher, P. M., et al.(2018). Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. Human molecular genetics, 27(20):3641–3649.23
- Nam, K., Kim, J., and Lee, S. (2022). Genome-wide study on 72,298 individuals in korean biobank data for 76 traits. Cell Genomics, 2(10):100189.
- Ferń andez-Rhodes, L., Graff, M., Buchanan, V. L., Justice, A. E., Highland, H. M., Guo, X., Zhu, W., Chen, H.-H., Young, K. L., Adhikari, K., et al. (2022). Ancestral diversity improves discovery and fine-mapping of genetic loci for anthropometric traits—the hispanic/latino anthropometry consortium. Human Genetics and Genomics Advances, 3(2):100099.
- Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., et al. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. Nature genetics, 53(10):1415–1424.