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

by

Guojun Ma
B.Sc., University of Alberta, 2020

Supervisory Committee

Dr. Yu-Ting Chen, (Co-)Supervisor
(Department of Mathematics and Statistics)

Dr. Xuekui Zhang, (Co-)Supervisor
(Department of Mathematics and Statistics)

ABSTRACT

Meta-analysis is a statistical approach that combines the results from multiple studies on the same scientific problem. in order to identify the overall effect or trend. This approach allows researchers to draw more robust conclusions compared to individual studies. This paper provides a concise review of methodology and software tools for conducting meta-analysis. Additionally, this paper presents a meta-analysis investigating the genetic associations with body mass index (BMI) across diverse ethnic populations, including European, Asian, and Latino groups. By analyzing a wide range of published GWAS studies involving various ethnicities, the aim is to identify shared genetic variants associated with BMI across populations, as well as potential population-specific markers. The results demonstrate that while certain genes, such as the FTO gene, consistently exhibit significant associations across ethnicities, there are also variations between populations. The implications of these meta-analysis findings are discussed, along with notable methodological considerations that arise in the process.

Table of Contents

Supervisory Committee i

Abstract ii

Table of Contents iii

Acknowledgements iv

Chapter 1 Introduction 1

Chapter 2 Review 4

2.1 Methods 4

2.1.1 Fixed-effect meta-analysis 5

2.1.2 Random-effect meta-analysis 7

2.1.3 Measuring heterogeneity 7

2.2 Software 9

Chapter 3 Meta-analysis of GWAS of body mass index 11

3.1 Cohorts information 11

3.2 Results 12

3.2.1 Significant loci 12

3.2.2 Heterogeneity analysis 15

3.3 Discussion 18

Bibliography 21

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my parents for their unwavering support and encouragement throughout my educational journey. I would also like to extend my sincere appreciation to the instructors in the math department for their exceptional dedication and commitment to imparting knowledge. I am thankful to my supervisors, Dr. Yu-ting Chen and Dr. Xuekui Zhang, for their guidance throughout the course of my project. A special word of thanks goes to Yupeng Zhao for collaborating with me on this project.

Chapter 1

Introduction

The Genome-wide association study (GWAS) is an observational research method that investigates the relationship between genetic variations and susceptibility to diseases or traits across the entire genome. This approach primarily focuses on analyzing variations in single nucleotide polymorphisms (SNPs), which are the most prevalent type of genetic variations in the genome. Since its initial publication in 2005 by the Wellcome Trust case control Consortium (WTCCC) [Klein et al., 2005], GWAS has gained substantial popularity within the scientific community. Over the years, researchers have successfully identified approximately $55,000$ unique loci in the genome associated with nearly $5,000$ diseases and traits [MacArthur et al., 2017]. The widespread adoption of GWAS has been made possible by advancements in sequencing technology, leading to a significant reduction in the cost of sequencing the entire genome. The result of GWAS not only assists researchers to gain insight into phenotype's underlying biology, but it can also aid medical practitioners in evaluating the likelihood of developing disease for patients and offering personalized treatment.

In the realm of scientific research, it is not uncommon to encounter discrepancies when multiple studies investigate the same problem using distinct experimental designs and methodologies, often leading to divergent and conflicting conclusions. To address this issue, researchers often perform systematic reviews. This involves gathering all available research related to a specific subject and methodology, and assessing and interpreting their findings. In the process of systematic reviews, the researchers

often employed meta-analysis, which is a statistical method for integrating numerical results from multiple studies. By synthesizing past evidence, systematic review and meta-analysis provide a more comprehensive and objective summary of the available evidence pertaining to a specific research question. Consequently, it is widely regarded as the strongest level of evidence within the field of evidence-based medicine literature [Herrera Ortiz et al., 2022].

When using meta-analysis for GWAS study, it offer several benefits. First, it increases the statistical power of association testing, which helps to recover signals that might be missed by single studies due to the small sample size. Additionally, meta-analysis enhances the precision and robustness of research findings. Furthermore, it facilitates the examination of cross-ancestry replicability and variability of genetic effects, providing valuable insights into the genetic architecture underlying the research question. However, it is crucial to approach meta-analysis with careful consideration, as careless or inadequate execution can yield misleading results. One prominent concern is publication bias, whereby the effect estimates may be inflated due to the selective publication of significant findings while negative or inconclusive results are often left unpublished. Moreover, heterogeneity among studies can significantly impact the outcomes of a meta-analysis, stemming from factors such as variations in population structures, genotyping platforms, environmental conditions, or phenotype measurements.

Scientists have been interested in investigating genetic associations with body mass index (BMI), a widely used measure of body fat based on an individual's height and weight. Numerous studies have leveraged GWAS to explore the genetic underpinnings of BMI. For example, the study by [Speliotes et al., 2010] identified 32 loci associated with BMI using data from 46 studies and a sample size of up to $123,865$ individuals. Building upon these findings, [Locke et al., 2015] expanded the scope by aggregating data from 80 GWAS studies and 34 Metabochip studies, ultimately identifying 97 loci associated with BMI. Furthermore, [Yengo et al., 2018] conducted a meta-analysis with a large sample size of approximately $700,000$ individuals, leading to the discovery of over 900 loci associated with BMI. Notably, a trend observed in these studies is

that larger sample sizes contribute to the identification of more significant loci, which encourages researchers to use an even larger sample size in future studies.

It is important to acknowledge a crucial limitation in the existing body of literature: the majority of studies with large sample sizes predominantly consist of participants of white European descent, thus overlooking a significant portion of the world's population. For instance, a research paper by [Nam et al., 2022] explores the genome-wide associations with BMI using data from the Japan Biobank.

The outline of this project is as follows: Chapter 2 provides an overview of the fundamental methods of meta-analysis, along with an examination of the available software commonly used. Chapter 3 presents the results of the meta-analysis conducted as part of this project. Specifically, our meta-analysis incorporates recent studies on the Japanese, Korean, Taiwanese, and Latino/Hispanic populations. Our objective is to discover novel genetic variations in the human genome that influence body mass index (BMI). Additionally, we aim to compare the effects of these genetic variations among different ethnic groups.

Chapter 2

Review

2.1 Methods

There are generally two types of meta-analysis that are commonly performed in practice. The first type is aggregate-based meta-analysis, which combines summary data such as means, standard deviations, and sample sizes from individual studies to estimate the overall effect. The second type is patient-based meta-analysis, which pools individual-level data from multiple studies instead of just the summary data. In general, patient-based meta-analysis is often more time-consuming and cumbersome compared to aggregate-based meta-analysis, since it requires the researchers to access, manage and analyze large sets of data from different sources. Furthermore, the paper showed that these two types of meta-analysis are similar in terms of statistical power [Lin and Zeng, 2010].

Meta-analysis commonly follows the following basic principles:

- In the first stage, calculate the summary statistics from each study, which describe the effect estimate in a consistent way. Depending on the nature of the study, the data type can be continuous, binary, among others.

- In the second stage, calculate the combined effect estimate as the weighted average of the effects estimated in each study:

$$\hat{\theta} = \frac{\sum Y_i W_i}{\sum W_i},$$

where Y_i of the effect estimated in the i-th study, W_i is the weight given to the i-th study, and the summation is across all studies.

- Consider whether the true effect is the same or varies across different studies. If it is assumed the true effect is the same, then performed a fixed-effect meta-analysis. Alternatively, perform a random-effects meta-analysis in which the estimated effects will follow a certain distribution.

- Assess whether there is heterogeneity among the results of the separate studies. Test whether the variations are due to statistical error or a true difference.

Let us assume there are a total of K independent studies. For each study, we observed the following effects on each study:

$$Y_i = \theta_i + e_i, \tag{2.1}$$

where Y_i denotes the observed effect in the i-th study, θ_i denotes the corresponding unknown true effect and e_i denotes the sampling error for i-th study with variance s_i^2. While The fixed-effect model treats θ_i to be the same across all studies, the random-effect model assumes the true effect θ_i as a random variable with mean μ and variance σ^2.

2.1.1 Fixed-effect meta-analysis

The most common and simple version of the meta-analysis procedure is referred to as the inverse-variance method. It can be applied to different types of effect measures, such as risk ratios, odds ratios, or mean differences. This method assigns the weight to each study to be the inverse of the variance of the effect estimate. Studies with larger sample sizes typically have smaller standard errors, resulting in a greater weight being assigned. It is commonly used for the fixed-effect model where the true effect is common to all studies. The weighted average is

$$\hat{\theta} = \frac{\sum Y_i(1/s_i^2)}{\sum(1/s_i^2)},$$

where the sum is over all studies. The standard error of the pooled effect estimate is

$$SE(\hat{\theta}) = \sqrt{\frac{1}{\sum(1/s_i^2)}}.$$

The confidence interval is given as $\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta})$, where $z_{\alpha/2}$ is the critical value from a standard normal distribution corresponding to a given significance level α.

In certain cases, the effect estimate and standard error may not be available in the summary statistics of an individual study. Instead, only the sample size and direction of effect estimated are provided. In such scenarios, the Z-score method can be used as an alternative approach, which is implemented in the software METAL[Willer et al., 2010]. Suppose there are multiple studies with each sample size of N_i, the direction of effect for each study Δ_i and p-value P_i are given. The Z-score is

$$Z_i = \Phi^{-1}(P_i/2) \times sign(\Delta_i), \tag{2.2}$$

where $\Phi(\cdot)$ denotes the cumulative normal distribution. The smaller p-value is assigned to the larger Z-score and vice versa. The overall Z-Score is combined across samples in a weighted sum, with weights proportional to the square root of the sample size for each study:

$$Z = \frac{\sum_i Z_i \sqrt{N_i}}{\sqrt{\sum_i N_i}}, \tag{2.3}$$

also, the overall p-value is $P = 2\Phi(-|Z|)$.

There are various methods that can be used to conduct fixed-effect meta-analyses. The Mantel-Haenszel method [Mantel and Haenszel, 1959] is best suited for binary effect measures and sparse data, and uses a different weighting scheme depending on the type of data being analyzed. Another method is the Peto method [Yusuf et al., 1985], which is appropriate for binary data with rare events.

2.1.2 Random-effect meta-analysis

The inverse variance method can be biased or misleading when there is heterogeneity among the studies since it does not account for the variation in the true effects across studies. In the presence of heterogeneity, the true effects θ_i are assumed to be random variables with mean μ and variance σ^2. In order to obtain an estimate of the true effect, it is common to employ a two-step approach. First, obtain an estimate of the variance as $\hat{\sigma}^2$. Then, estimate the mean of θ_i as

$$\hat{\mu} = \sum \hat{w}_i Y_i / \sum \hat{w}_i,$$

where $\hat{w}_i = (\hat{\sigma}^2 + s_i^2)^{-1}$.

There are many methods for estimating the heterogeneity variance, which is compared and studied in the paper [Langan et al., 2019]. The method proposed by Dersimonian and Laird [DerSimonian and Laird, 1986] is most commonly used, and it is available in most software packages for meta-analysis. The estimator is defined as

$$\hat{\sigma}^2_{DL} = \max\left\{0, \frac{\sum_{i=1}^{k}(1/\hat{s}_i^2)(\hat{\theta}_i - \bar{\theta})^2 - (k-1)}{\sum_{i=1}^{k}(1/\hat{s}_i^2) - \frac{\sum_{i=1}^{k}(1/\hat{s}_i^2)^2}{\sum_{i=1}^{k}(1/\hat{s}_i^2)}}\right\},$$

where $\hat{s}_i$ is the estimated standard error for study i, $\hat{\theta}_i$ is the effect estimated for study i and $\bar{\theta}$ is the mean effect estimate. There are several other estimators proposed in the literature, such as Hartung-Makambi estimator [Hartung and Makambi, 2003], which is a correction of Dersimonian and Laird method so that the estimator is always non-zero.

2.1.3 Measuring heterogeneity

During the process of conducting a meta-analysis, it is frequently observed that there is heterogeneity among the effect estimates obtained from different groups. This heterogeneity can be due to a variety of factors, including differences in participant characteristics or variations in study design. For instance, when examining the effectiveness

of the COVID-19 vaccine, it may be found that the vaccine has a more pronounced impact on reducing mortality rates among older individuals compared to younger ones. Consequently, researchers undertaking a meta-analysis must take into account these variations in order to derive meaningful conclusions. Furthermore, investigating the source of heterogeneity holds scientific significance as it provides researchers with a more comprehensive understanding of the original problem.

Various methods can be utilized to measure the degree of heterogeneity in meta-analysis. One such method is the forest plot, which presents the effect estimate, standard error, and confidence interval for each study. If the confidence intervals of individual studies do not overlap, it indicates the presence of heterogeneity. Additionally, statistical tests like the χ^2 test statistic can be used to evaluate heterogeneity, which is defined as

$$Q = \sum_{i=1}^{k} \frac{(\theta_i - \hat{\theta})^2}{s_i^2},$$

where k is the number of studies, θ_i is the effect estimate for study i, $\hat{\theta}$ is the inverse-variance weighted average, and s_i is the standard error for study i. Under the null hypothesis that there is no heterogeneity, Q follow the χ^2 distribution with $k-1$ degrees of freedom. The p-value is can be obtained referring χ^2 distribution.

One way to measure the degree of heterogeneity is by using the I -squared statistic. It is defined as:

$$I^2 = \frac{Q - df}{Q} \times 100\%,$$

where Q is the χ^2 statistics and df is the degrees of freedom. I^2 describes the percentage of the variability in effect estimates that is due to heterogeneity. According to the Cochrane Handbook [Higgins et al., 2019], some rough guidelines for interpreting I-squared are:

- 0 % to 40 % : might not be important;
- 30 % to 60 % : moderate heterogeneity;
- 50 % to 90 % : substantial heterogeneity;

- 75 % to 100 % : Considerable heterogeneity.

When there is a significant amount of variation in a meta-analysis, researchers use different strategies to identify and explore the underlying sources of this variability. These strategies include meta-regression, subgroup analysis, or using a random-effects model. Meta-regression involves adding covariates to the meta-analysis model, such as age, gender, or ethnicity of the samples. Subgroup analysis is another valuable approach that compares the effect estimate within different subgroups. Alternatively, random-effect meta-analysis incorporates heterogeneity by modelling the true effect as a probability distribution.

2.2 Software

There are numerous software packages available for conducting meta-analysis. Some popular commercial options include MetaWin, Comprehensive Meta-analysis (CMA), and Review Manager 5 (RevMan 5), which are user-friendly but require a subscription to use. Alternatively, one can conduct meta-analysis using Microsoft Excel with add-ins such as MIX or MetaEasy. There are various packages in the R program - a paper by [Polanin et al., 2017] reviewed and compared 63 R packages designed for meta-analysis. Among these packages are general ones like meta, metafor, and rmeta, which provide the necessary functions to conduct a basic meta-analysis. For instance, metafor allows the user to calculate effect sizes, plot synthesis results, handle missing data, perform sensitivity analyses, and assess publication biases. Additionally, there are specialized packages designed for specific scientific disciplines, such as epiR, which is tailored for epidemiological data.

One unique issue in GWAS meta-analysis is the inconsistent coding of SNPs across different datasets, commonly known as the 'strand' issue. For instance, if a SNP has alleles A and T, one study may code A as the reference allele while another study may code T as the reference allele. This inconsistency can result in the reversal of the effect directions of the SNP in the meta-analysis. Some meta-analysis software have the capacity to automatically remove or correct mislabeled SNPs.

Another challenge is population stratification, which arises from differences in allele frequencies between subpopulations within a study or across different studies. This can introduce confounding factors in the association between SNPs and traits. For example, in a GWAS examining the association between a SNP and hypertension, if the study population includes both Europeans and Africans and the SNP has a higher frequency in Africans, a positive association between the SNP and hypertension may be observed, even if the SNP has no causal effect. In such cases, the ethnicity of the individuals can act as a confounding variable. Several methods, such as genomic control or principal component analysis, have been developed to address this issue and are often implemented in specialized software tools.

Furthermore, GWAS data files are commonly available in various formats and are often substantial in size. Therefore, an ideal software tool for GWAS should possess the capability to handle multiple file formats, efficient memory management, and fast computing capabilities. Table 2.1 provides an overview of some commonly used software packages along with their respective features.

Table 2.1: Common software used for GWAS meta-analysis

Software package	**METAL**	**GWAMA**	**PLINK**	**GWAR**
Fix-effect analysis	✓	✓	✓	✓
Random-effect analysis		✓	✓	✓
robust statistics				✓
File format and separator versatility	✓	✓		✓
Resolve label mismatch	✓	✓		
Genomic control correction	✓	✓		

The METAL [Willer et al., 2010] software is commonly used to perform meta-analysis. The METAL software operates in a command-line environment and can be run on Windows, Mac, and most Linux systems. It has efficient memory management and is relatively fast. It can handle various file formats and delimiters with ease. The software implements two types of meta-analysis: an inverse variance method and a weighted score method based on the sample size, p-value and direction of effect in each study. However, it does not incorporate a random-effect meta-analysis feature, nor does it generate graphs to visualize the result. It is common to complement the used for METAL with the R packages mention above.

Chapter 3

Meta-analysis of GWAS of body mass index

3.1 Cohorts information

The GWAS Catalog is a comprehensive database that compiles data from previously published genome-wide association studies. We conducted an extensive search of journal papers using the GWAS Catalog website and identified studies that included summary statistics of genome-wide association with body mass index (BMI). These summary statistics store the association level of single nucleotide polymorphisms (SNPs), which are the specific locations in the human genome that show the most variations in a population. For example, the study [Yengo et al., 2018] reports that a specific SNP located at chromosome 2 and base pair position 3221999 was found to cause an average increase of BMI of $0.021(s.e. \pm 0.01)$kg for individuals with an allele of Adenine(A) in this location, with a p-value of 5×10^{-10} indicating that the estimated effect is different from zero.

Our meta-analysis comprises several studies conducted from 2015 to 2022, and we have summarized the cohorts in Table 3.1. The largest cohort is the UK Biobank(UKB), which is a biomedical database that holds data on over $500,000$ European-descent individuals from the United Kingdom, including detailed health information, genetic data, and lifestyle factors. The UKB is a unique resource due to its size and scope and has

already contributed to many important discoveries. Similar biomedical databases have also been established in many other countries, such as Taiwan biobank, Korea biobank, and Japan biobank. In total, the sample size of the meta-analysis exceeds one million and comprises individuals from various ethnic backgrounds.

Table 3.1: Information of cohorts

Cohort name	**Sample size**	**Ethnicity**	**Reference**
UK Biobank	456,426	White European	[Yengo et al., 2018]
GIANT	322,154	White European	[Locke et al., 2015]
Korea biobank	72,298	East Asian	[Nam et al., 2022]
Taiwan Biobank	21,930	East Asian	[Wong et al., 2022]
Japan biobank	179,000	East Asian	[Sakaue et al., 2021]
HISLA	56,161	Hispanic/Latino	[Fernández-Rhodes et al., 2022]

3.2 Results

3.2.1 Significant loci

For the cohorts listed in the table 3.1, we conducted a fixed-effect inverse-variance weighted meta-analysis using the METAL software. We consider a subset of approximately 2.3 million SNPs showing consistent alleles with UKB and GIANT cohorts. The results of the meta-analysis can be visualized with the Manhattan plot 3.1, which shows the association level of SNPs with BMI across the entire genome. The plot annotates the genes that are nearest to the SNPs with the strongest association. The plot suggests a polygenicity phenomenon, in which multiple regions in the genome are significantly associated with BMI. Figure 3.5 displays the regional association plots for the genes AL136114.1 and THEM18, which shows that multiple SNPs in the region have a tendency to correlate and exhibit similar significant levels. We also evaluate the presence and degree of bias using the Q-Q plot 3.2, which indicates a minor bias of observed p-value causes by population stratification.

We used a strict significant threshold 5×10^{-8} for the p-value to address the issue of false positives, which is common in statistical genetic research due to the large number of SNPs involved. Out of the approximately 2.3 million SNPs analyzed, we identified

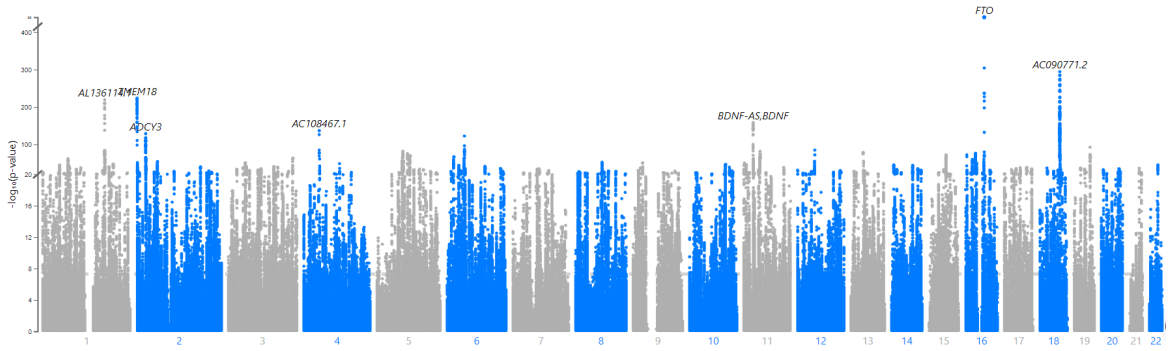

Figure 3.1: The Manhattan plot shows the significant level across the whole genome. The x-axis denotes the chromosome and location of SNPs. The y-axis denotes the p-value in a logarithmic scale of 10. The dashed line indicates the significant threshold of the p-value. The genes with the most significant effects are annotated.

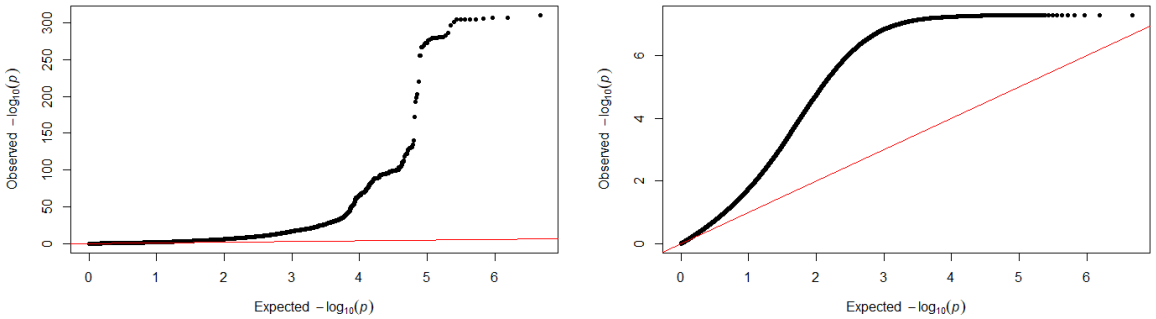

Figure 3.2: Q-Q plot compares the observed and expected distributions of p-values. The x-axis shows the expected p-values under the null hypothesis of no association, and the y-axis shows the observed p-values for each SNP. The p-values are plotted on a log 10 scale to highlight the small values. The left-hand side shows the Q-Q plot of the Meta-analysis results, while the right-hand side shows the Q-Q plot after removing all significant SNPs with $p < 5 \times 10^{-8}$. The sample quantile on the right-hand side is slightly above the diagonal line, indicating a minor negative bias of p-value.

61,507 SNPs that were statistically significant. This corresponds to a total of 1,966 in different loci, where each locus was defined as a window of 500 Kb. For comparison, a previous GWAS of BMI [Yengo et al., 2018] identified a total of 41,103 significant associated SNPs correspond to 1,239 loci. In total, we identified 27,616 new significant associated SNPs.

Generally, many non-causal variants are significantly associated with a trait of interest due to linkage disequilibrium. Those significant SNPs are clustered in loci,

which are sets of correlated variants that all show a significant association with the trait of interest. To identify causal variants, further analysis is required, such as conditional association analysis using GCTA-COJO software [Yang et al., 2012]. However, we do not have access to the necessary information, such as the sample size of variants, to perform this analysis. As an alternative, we use a simple approach of identifying the likely causal variant by selecting the SNP with the lowest p-value among loci with more than 10 significant SNPs. In this way, we have identified 774 different loci, in which the location can be visualized in a bar plot 3.3. In particular, we discovered the highest number of new loci in chromosome 2. Further, the largest density of association SNPs was observed on chromosome 2 near the genes NBAS and DDX1, where 193 are clustered within 500kB of each other.

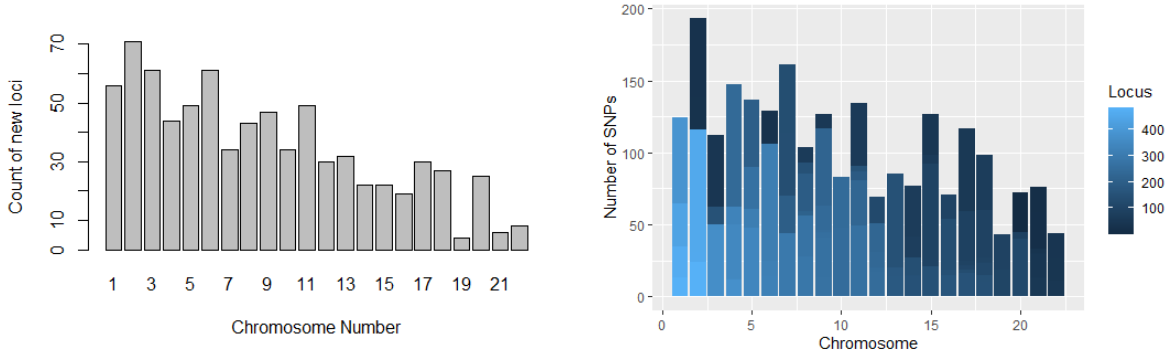

Figure 3.3: The left-hand side displays the distribution of newly discovered loci. The right-hand side is a bar plot displaying the number of newly discovered SNPs in various loci. The horizontal axis represents the chromosome, while the vertical axis represents the number. The colour gradient is used to differentiate between different loci within a chromosome.

Table 3.4 presents the top 10 new loci with the most significant association. The result indicates that the A/G alleles at chromosome 6 position 20655110 lead to an average increase of 0.0249 in BMI. The nearest gene to this particular SNP is CDKAL1. The effect estimate direction is consistent across all studies, but the I^2 and χ^2 test statistics suggest significant heterogeneity in the effect estimated across studies. The regional plot in 3.5 shows the association level in the regions around the markers rs9368216 and rs10515239. The top plot indicates that some of the significant associated SNPs lie in the intergenic area, while the bottom plot shows that they lie in the

gene area called CDKAL1.

CHR	POS	MarkerName	Nearest_gene	Allele1	Allele2	Effect	StdErr	P.value	Direction	HetISq	HetDf	HetChiSq	HetPVal
5	95852769	rs10515239	CDKAL1	c	g	0.0308	0.0020	5.589e-54	+++++	35.0	4	6.150	1.882e-01
19	46221726	rs11881883	RPL12P41	a	g	-0.0292	0.0019	2.970e-55	---+-	90.7	4	43.158	9.594e-09
16	20258432	rs12597682	LINC01554	a	c	-0.0281	0.0021	7.664e-40	-----	77.3	4	17.594	1.481e-03
3	52738165	rs12635140	KCNQ1	t	c	-0.0146	0.0013	2.451e-30	-----	82.0	4	22.228	1.805e-04
11	27635319	rs1387144	SNRPEP3	a	c	0.0157	0.0013	2.435e-34	+++++	93.0	4	57.423	1.009e-11
6	34183057	rs1592269	CDKN2B-AS1	a	g	-0.0285	0.0023	3.386e-36	-----	40.8	4	6.755	1.494e-01
11	2839751	rs2237892	KRT18P9	t	c	0.0258	0.0017	5.058e-51	+++++	89.2	4	36.890	1.898e-07
10	122913475	rs7098433	BDNF-AS	t	c	-0.0227	0.0020	9.006e-30	-----	67.2	4	12.206	1.588e-02
16	53833605	rs7204609	FTO	t	c	0.0223	0.0019	1.344e-30	+++++	0.0	4	2.332	6.750e-01
6	20655110	rs9368216	NEK4	a	g	0.0249	0.0015	1.260e-59	+++++	91.3	4	46.106	2.341e-09

Figure 3.4: The following table shows the 10 new SNPs with the most significant estimated effect. The table includes information on the SNP marker label, chromosome and position of the SNP, the nearest gene, alleles, estimated estimate, standard error, and p-value. The "hetISq" column provides I^2 statistics, which measure heterogeneity on a scale of $0-100\%$. Additionally, the "HetChiSq", "HetDf", and "HetPval" columns indicate the chi-squared statistics, degree of freedom, and p-value of the test statistics, respectively.

We utilized the UCSC Genome Browser gateway to conduct research on the information of 10 genes with the most significant association level. The provided table 3.2 presents details regarding the type and function of these genes. CDKAL is a gene that codes for proteins and is expressed in pancreatic islets. Previous studies have linked this gene to susceptibility to type 2 diabetes, which is not surprising given that BMI is also associated with diabetes. In addition, there exist genes located in the intergenic regions that do not play a role in protein coding, but rather serve as regulators and facilitators.

3.2.2 Heterogeneity analysis

We discovered that the new significant loci have significant heterogeneity of effect between various groups. The forest plot 3.6 displays the effect estimate from the different studies. The plot indicates that the effect estimates for SNP marker rs9368219 are positive for the East Asian population, but negative for the European and Hispanic populations. Conversely, rs1861866 appears to exhibit the opposite effects. For SNP markers rs979614 and rs7103873, the effect estimates are not statistically significant for European and Hispanic populations, but demonstrate a negative effect on the Asian

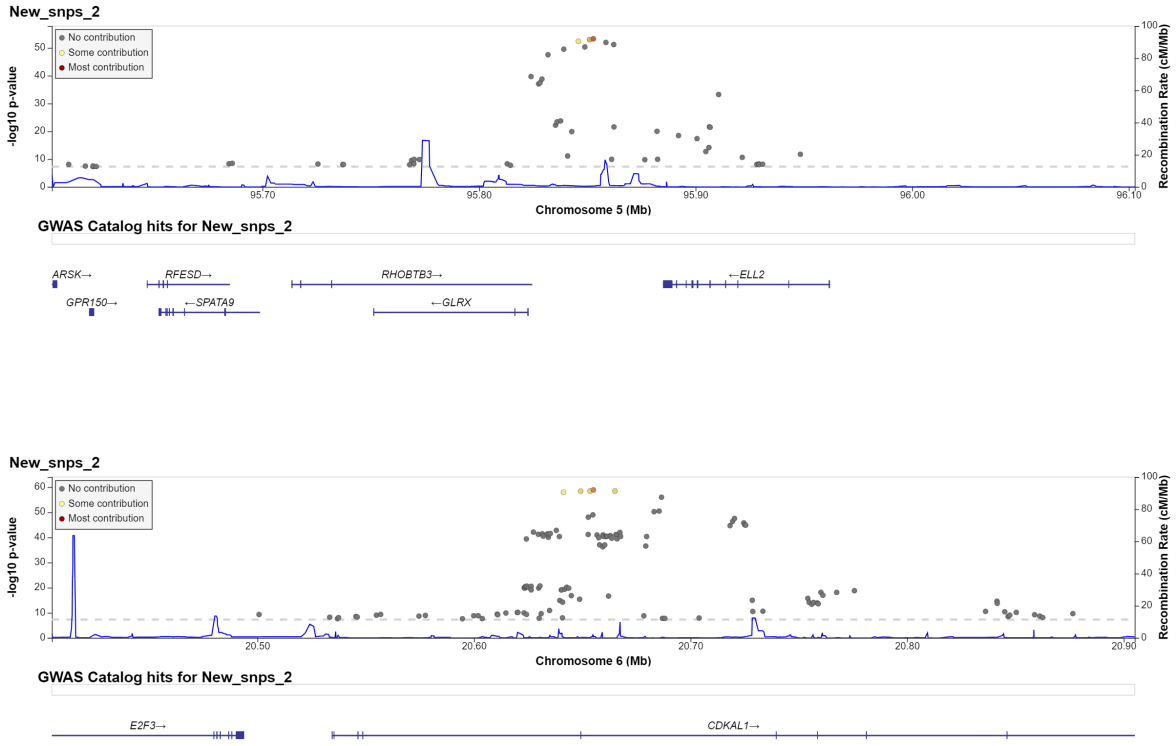

Figure 3.5: A regional plot is shown for the genes ELL2 in chromosome 5 and CDKAL1 in chromosome 6. The x-axis indicates the position of the SNP(measure in Mb), while the y-axis represents the p-value in a logarithmic scale of 10. The colour gradient highlights the SNPs that are likely the causal variants. The bottom of each plot indicates the location of the protein-encoding genes.

population.

We also conducted separate meta-analyses for each ethnic group. The study identified genome-wide significant SNPs ($p < 5 \times 10^{-8}$) in the European, East Asian, and Latino/Hispanic groups, which are summarized in table 3.3. Note that meta-analysis enhances the power of Genome-Wide Association Studies (GWAS). As a result, the trans-ancestry meta-analysis detects a greater number of significant SNPs ($61,507$) compared to the sum of its subgroups. To compare the results across the ethnic groups, a Manhattan plot was used to display the genetic variations of different populations 3.7. The plot indicates that the FTO gene is the most common variant with the highest level of significance for all populations. Additionally, genes such as RSL24D1P11,

Gene	Type	Tissue specificity	Disease association
CDKAL1	protein coding	pancreatic islets	type 2 diabetes
RPL12P41	pseudogene	-	-
LINC01554	intergenic area	-	-
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

Table 3.2: Functional summary of discovered genes

AL136114.1, THEM18, and others are also significantly associated with BMI across all populations. The study also identified 492 significant loci that are uniquely in the Asian population and 2 significant loci that are uniquely in the Latino/Hispanic population, which is summarised in table 3.4.

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%)

Table 3.3: Summary of results from within-ancestry and trans-ancestry GWAS meta-analyses

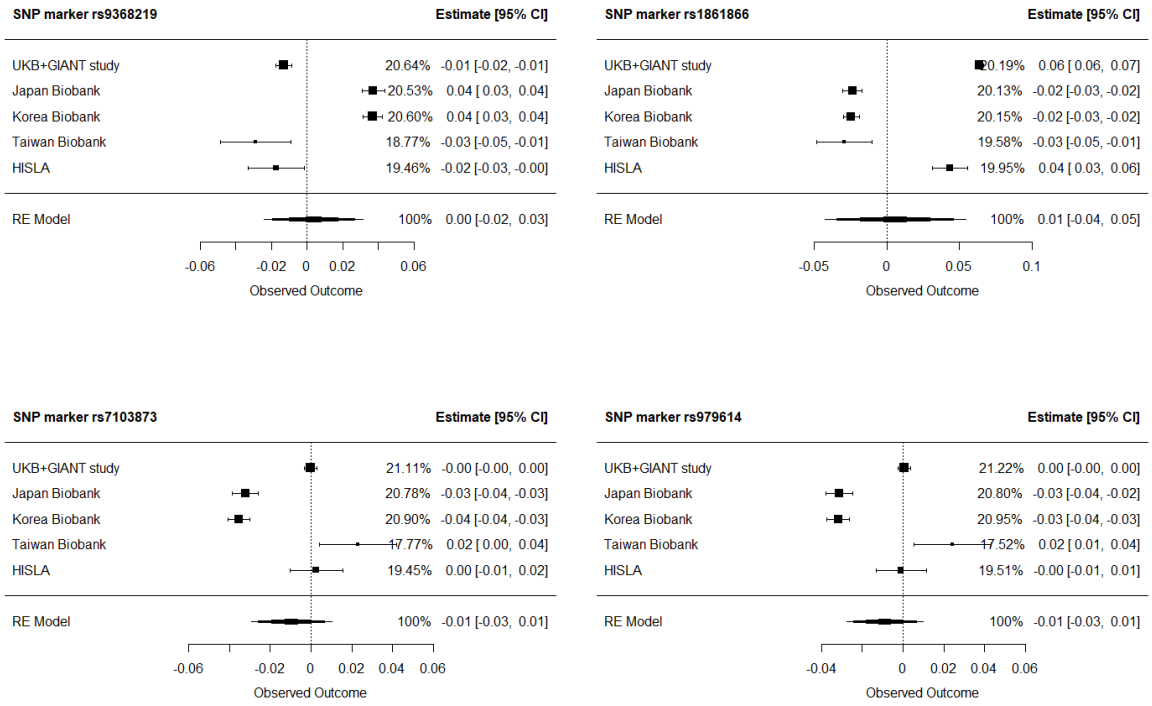

Figure 3.6: The forest plots display the effect estimate and confidence interval of various studies

Chromosome	Lead SNP	p-value	number of SNPs in a locus
19	rs11671664	4.844^{-84}	81
5	rs261967	4.016^{-63}	86
11	rs2237892	1.123^{-54}	24
9	rs10965250	7.049^{-50}	6
4	rs1996023	4.439^{-45}	14
⋮	⋮	⋮	⋮
1	rs545608	1.924^{-10}	10
16	rs1558902	3.391^{-32}	51

Table 3.4: Unique discovery in the East Asian population(shown on the top) and Latino/Hispanic population(shown on the bottom)

3.3 Discussion

We have discovered numerous new loci that show significant association with BMI. Some of these loci are common across all ethnic groups, while some are unique to specific groups. To obtain a comprehensive map of genetic variants linked to human

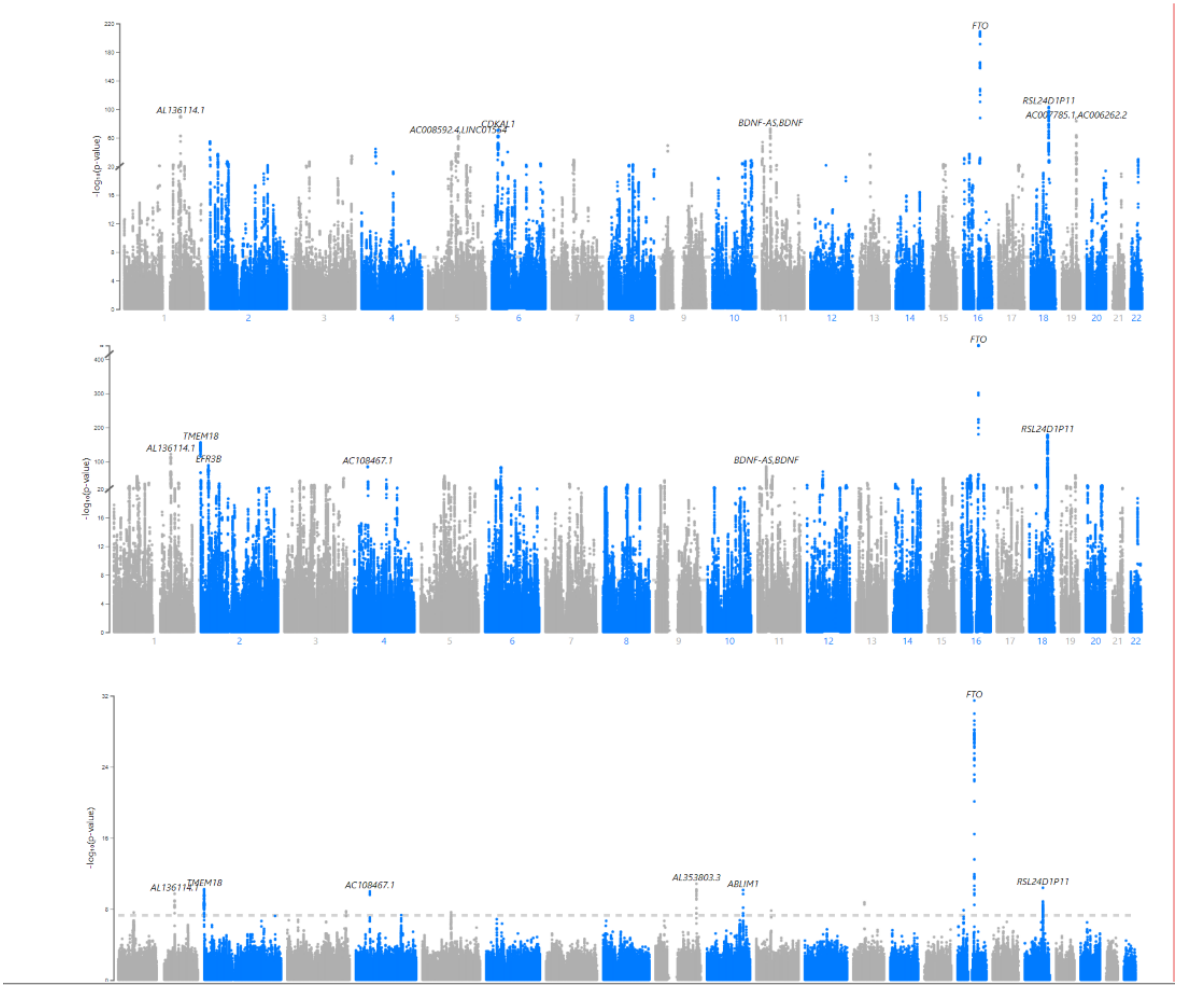

Figure 3.7: The Manhattan plot displays the genetic variations of different populations, with the Asian population at the top, the European population in the middle, and the Latino/Hispanic population at the bottom.

weight, we suggest future studies with even larger sample sizes and more diverse populations. Additionally, a potential avenue for future research is to explore the biological mechanisms of these significant SNPs. This could involve identifying the immediate effects of causal variants, such as whether they are responsible for protein-encoding or serve as enhancers, as well as examining the network effects that lead to changes in cellular and physiological function.

Our research has revealed a significant increase in the number of associated SNPs compared to previous studies that focused only on the European population [Yengo et al., 2018]. However, it is important to acknowledge the potential bias introduced by conduct-

ing a cross-ethnicity meta-analysis. To mitigate this, various methods such as genomic control (GC) correction, principal component analysis, or linkage disequilibrium score regression (LDSC) can be used. Nevertheless, none of these methods can perfectly correct the bias, especially with a large sample size. In fact, a current research area is finding a more effective method for addressing population stratification bias [Uffelmann et al., 2021].

Our findings suggest that BMI traits are not determined by a single genetic variant, but rather by multiple genetic variants, each of which has a small effect. This polygenic nature of BMI is also observed in other similar traits such as height, skin colour, and various diseases. Additionally, environmental factors and gene-environment interactions play a significant role in determining these traits. As a result, understanding the biological mechanisms of these variants and exploring potential therapeutic interventions is challenging - as in the previous research of GWAS with BMI, the result SNPs explain 6% of the variance in the population. Novel methods are required to address polygenicity and facilitate the translation of GWAS findings into biological insights.

Bibliography

[DerSimonian and Laird, 1986] DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188.

[Fernández-Rhodes et al., 2022] Fernández-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.

[Hartung and Makambi, 2003] Hartung, J. and Makambi, K. H. (2003). Reducing the number of unjustified significant results in meta-analysis. *Communications in Statistics-Simulation and Computation*, 32(4):1179–1190.

[Herrera Ortiz et al., 2022] Herrera Ortiz, A. F., Cadavid Camacho, E., Cubillos Rojas, J., Cadavid Camacho, T., Zoe Guevara, S., Tatiana Rincón Cuenca, N., Vásquez Perdomo, A., Del Castillo Herazo, V., and Giraldo Malo, R. (2022). A practical guide to perform a systematic literature review and meta-analysis. *Principles and Practice of Clinical Research*, 7(4):47–57.

[Higgins et al., 2019] Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

[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.

[Langan et al., 2019] Langan, D., Higgins, J. P., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W., and Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research synthesis methods*, 10(1):83–98.

[Lin and Zeng, 2010] Lin, D. and Zeng, D. (2010). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 34(1):60–66.

[Locke et al., 2015] Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206.

[MacArthur et al., 2017] 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.

[Mantel and Haenszel, 1959] Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4):719–748.

[Nam et al., 2022] 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.

[Polanin et al., 2017] Polanin, J. R., Hennessy, E. A., and Tanner-Smith, E. E. (2017). A review of meta-analysis packages in r. *Journal of Educational and Behavioral Statistics*, 42(2):206–242.

[Sakaue et al., 2021] 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.

[Speliotes et al., 2010] Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*, 42(11):937–948.

[Uffelmann et al., 2021] Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59.

[Willer et al., 2010] Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191.

[Wong et al., 2022] Wong, H. S.-C., Tsai, S.-Y., Chu, H.-W., Lin, M.-R., Lin, G.-H., Tai, Y.-T., Shen, C.-Y., and Chang, W.-C. (2022). Genome-wide association study identifies genetic risk loci for adiposity in a taiwanese population. *PLoS Genetics*, 18(1):e1009952.

[Yang et al., 2012] Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., of ANthropometric Traits (GIANT) Consortium, G. I., Replication, D. G., analysis (DIAGRAM) Consortium, M., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2012). Conditional and joint multiple-snp analysis of gwas summary

statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369–375.

[Yengo et al., 2018] 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.

[Yusuf et al., 1985] Yusuf, S., Peto, R., Lewis, J., Collins, R., and Sleight, P. (1985). Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in cardiovascular diseases*, 27(5):335–371.