Heliyon 10 (2024) e36023

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Genome-wide association study based on clustering by obesity-related variables uncovers a genetic architecture of obesity in the Japanese and the UK populations

Ippei Takahashi [a], Hisashi Ohseto [a], Fumihiko Ueno [a,b], Tomomi Oonuma [b], Akira Narita [b], Taku Obara [a,b,c], Mami Ishikuro [a,b], Keiko Murakami [a,b], Aoi Noda [a,b,c], Atsushi Hozawa [a,b], Junichi Sugawara [a,b], Gen Tamiya [a,b,d], Shinichi Kuriyama [a,b,e,*]

[a] *Graduate School of Medicine, Tohoku University, Sendai, Japan*
[b] *Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan*
[c] *Tohoku University Hospital, Sendai, Japan*
[d] *RIKEN Center for Advanced Intelligence Project, Tokyo, Japan*
[e] *International Research Institute of Disaster Science, Tohoku University, Sendai, Japan*

A B S T R A C T

Whether all obesity-related variants contribute to the onset of obesity or one or a few variants cause obesity in genetically heterogeneous populations remains obscure. Here, we investigated the genetic architecture of obesity by clustering the Japanese and British populations with obesity using obesity-related factors. In Step-1, we conducted a genome-wide association study (GWAS) with body mass index (BMI) as the outcome for eligible participants. In Step-2, we assigned participants with obesity (BMI $\geq$25 kg/m^2) to five clusters based on obesity-related factors. Subsequently, participants from each cluster and those with a BMI <25 kg/m^2 were combined. A GWAS was conducted for each cluster.

Several previously identified obesity-related genes were verified in Step-1. Of the genes detected in Step-1, unique obesity-related genes were detected separately for each cluster in Step-2. Our novel findings suggest that a smaller sample size with increased homogeneity may provide insights into the genetic architecture of obesity.

1. Introduction

Obesity is a serious global medical and economic issue that represents a major risk factor for many lifestyle-related diseases, such as diabetes, hyperlipidemia, and hypertension [1,2]. The global proportion of individuals with a body mass index (BMI) $\geq$ 25 kg/m^2 is reportedly 36.9 % and 38.0 % for men and women, respectively [3]. The pathogenesis of obesity is complex and includes regulation of calorie utilization, appetite, and physical activity, including health care availability, socioeconomic status, and underlying genetic and environmental factors [4,5].

Heritability of BMI has been extensively reported. For instance, in twin studies, BMI heritability ranged from 30 % to 90 % [6–8], whereas in genome-wide association studies (GWASs), it was estimated to be 20–30 % [9–12]. Only approximately 6 % has been reported based on genome-wide significant loci [9–11]. Although GWASs using BMI as an outcome have identified over 700 associated loci [9–15], whether they all contribute to obesity development via the same pathway remains obscure. The association of these

* Corresponding author. Graduate School of Medicine, Tohoku University, Sendai, Japan.
E-mail address: shinichi.kuriyama.e6@tohoku.ac.jp (S. Kuriyama).

https://doi.org/10.1016/j.heliyon.2024.e36023
Received 8 June 2023; Received in revised form 6 August 2024; Accepted 8 August 2024
Available online 9 August 2024

genetic variants with obesity may be elucidated by a polygenic model, wherein the effects of each variant are weak yet contribute to the onset of obesity [16]. Hence, if the genetic architecture of obesity could be elucidated by a polygenic model, we would expect larger sample sizes to correspond to more identified signals, whereas smaller sample sizes would correspond to fewer signals (Fig. S1). Moreover, within a genetically heterogeneous population of obesity, if a few variants exhibit a relatively strong influence leading to obesity in a portion of the subtypes included therein, then dividing the population with obesity into homogeneous groups can detect unique genes in each population, even with a reduced sample size. However, to the best of our knowledge, no GWASs has been conducted to divide people with obesity into more homogeneous populations. In a previous study, Traylor et al. demonstrated that categorizing patients with a complex disease into more homogeneous subgroups offered more insight into hidden heritability in a simulation study [17]. Thus, clustering algorithms for machine learning could reveal novel and more genetically homogeneous clusters.

Despite these circumstances, no studies have examined the genetic architecture of obesity by dividing obese individuals into clusters using machine learning techniques. Herein, we investigated the genetic architecture of obesity by dividing individuals with obesity into clusters using diverse obesity-related factors and machine learning techniques and conducting GWAS on each cluster (cluster-based GWAS) [18,19].

2. Materials and methods

2.1. Resource availability

The data that support the findings of this study are available from the TMM biobank. However, restrictions apply to the availability of these data, which were accessed under the license of the current study. Hence, they are not publicly available. Data are available from the authors upon reasonable request and with permission from the TMM Biobank. All inquiries regarding access to data should be addressed to TMM Biobank at dist@megabank.tohoku.ac.jp.

2.2. Experimental model and study participant details

2.2.1. Population

This study was conducted in accordance with the guidelines of the Declaration of Helsinki [20]. The protocol was reviewed and approved by the Institutional Review Board of the TMM Organization. The main study used data from cohort studies conducted by the TMM Birth and Three-Generation Cohort Study (BirThree Cohort Study) and the TMM Community-Based Cohort Study (CommCohort Study) [21–23]. The BirThree Cohort and CommCohort Study were conducted in the Miyagi and Iwate Prefectures, Japan. Details of the BirThree Cohort and CommCohort Study have been described previously [22,23]. The BirThree Cohort Study was a birth- and three-generation cohort study. Pregnant women were registered between July 2013 and March 2017 [22]. Additionally, the family of the pregnant women and their partners (biological father) were recruited (including maternal and paternal grandparents, siblings of the fetus, and their relatives) [22]. Among the BirThree Cohort Study participants, mothers (n = 22,493), fathers (n = 8,823), and grandparents (n = 8058) of the fetus were included in this study. The TMM CommCohort study is a community-based prospective cohort study that includes men and women aged >20 years living in Miyagi Prefecture, Northeastern Japan [23]. The Type 1 survey (n = 41,097) was conducted at specific municipal health check-up sites. The Type 2 survey (n = 13,855) was conducted at assessment

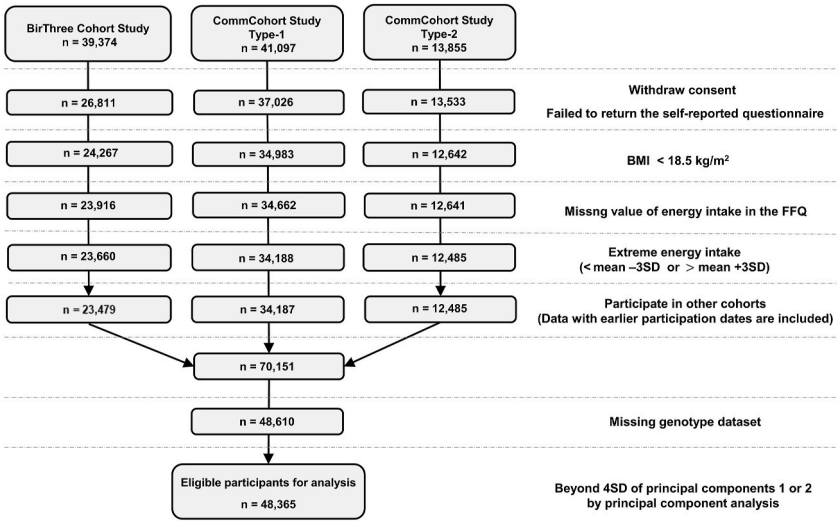

Fig. 1. Flow chart of exclusion criteria in this study
Participant data from each cohort are excluded based on these criteria.

centers [23].

Participant data were excluded based on the following criteria: withdrawal of consent, failure to return the self-reported questionnaire, BMI <18.5 kg/m^2, missing information on the food frequency questionnaire (FFQ), extreme energy intake (energy intake > mean ± 3 SD), and duplicate participation in both the BirThree Cohort and the CommCohort Study Type-1 (the data of participation at an earlier date were included). Data from eligible participants of the BirThree Cohort Study (n = 23,479), CommCohort Study Type-1 (n = 34,187), and CommCohort Study Type-2 (n = 12,485) were combined (n = 70,151). In the sub-analysis, a similar analysis was conducted using UKB data [24–26] to compare the results with those of the main study. The methods for analyzing the UKB data are described in the Supplementary Information.

2.2.2. Genotyping, imputation, and quality control

Cohort participants were genotyped using the Affymetrix Axiom Japonica Array (v2) in 19 batches, with 50 plates per batch. Details pertaining to the genotyping conducted in TMM have been described previously [27]. Following batch genotyping, samples with a call rate <0.95 or samples with unusually high IBD values compared to other samples were excluded. Additionally, variants with Hardy–Weinberg equilibrium P-values $<1.00 \times 10^{-5}$, minor allele frequency <0.01, or missing fraction >0.01 were excluded from each batch. A direct genotype dataset in the PLINK BED format was obtained by merging the genotype datasets of the 19 batches. A total of 21,541 participants with missing direct genotypic data were excluded. Principal component analysis (PCA) was conducted using the pca approx tool in PLINK 2.0 [28] on the direct genotype dataset. An additional 245 participants with >4 SD for principal components 1 or 2 were excluded. Finally, 48,365 participants (BirThree Cohort Study: n = 11,674; CommCohort Study Type-1: n = 27,745; and CommCohort Study Type-2: n = 8,946) were included in the analysis (Fig. 1). A plot of the participants (n = 48,365) according to principal components 1 and 2 using PCA is depicted in Fig. S2.

To prepare an imputed genotype dataset, pre-phasing was conducted using SHAPEIT2 [29], along with the Duo tool [30], which incorporates information on the relatedness between individuals to increase phasing accuracy. The phased genotypes were subsequently imputed with a cross-imputed panel of 3.5KJPNv2 [31] and 1KGP3 [32] using IMPUTE4 [26]. To create the cross-imputation panel for 3.5KJPNv2 [31] and 1KGP3 [32], the merge_ref_panels tool in IMPUTE2 was used [33]. Consequently, we obtained an imputed genotype dataset in the Oxford BGEN format (https://www.chg.ox.ac.uk/~gav/qctool_v1/#overview). For genotype imputation data, those with minor allele frequencies <0.01 and imputation information scores <0.8 were excluded. Finally, 9,868,333 single nucleotide polymorphisms (SNPs) were included in the GWASs.

2.3. Quantification and statistical analysis

2.3.1. Variables

The following variables related to obesity, which were collected from questionnaires completed by the participants at baseline for each cohort, were used for clustering: age, nutrient intake calculated from the FFQ based on frequency of food intake over the past year (energy, protein, fat, carbohydrate, sodium, potassium, calcium, magnesium, phosphorus, iron, zinc, copper, manganese, retinol equivalents, vitamins [C, D, K, B1, B2, B6, and B12], niacin, folate, pantothenic acid, cholesterol, dietary fiber, lycopene, α-carotene, β-carotene, and β-cryptoxanthin), frequency of leisure time physical activity (slow or fast walking, moderate exercise, and strenuous exercise; the choices included: no activity, more than once per month, 1–3 times per month, 1–2 times per week, 3–4 times per week, and almost every day), time typically spent in physical activity per day (strenuous work, walk, standing, and sitting) according to predetermined options (no time, <30 min, >30 min ≤60 min, >1 h ≤ 3 h, >3 h ≤ 5 h, >5 h ≤ 7 h, >7 h ≥ 9 h, >9 h ≤ 11 h, and >11 h) sleep duration (<5 h, >5 h ≤ 6 h, >6 h ≤ 7 h, >7 h ≤ 8 h, >8 h ≤ 9 h, and >9 h), difference between weight at age 20 years and current weight, smoking (smoked >100 cigarettes since birth; yes or no), alcohol consumption (>1 drink per month, quit, rarely, and unable to drink), psychological distress over the past month (total K6 score [Japanese version]) [34,35], and birth weight (unknown; >1,500 g ≤ 2,000 g; >2,000 g ≤ 2,500 g; >2,500 g ≤ 3,000 g; >3,000 g ≤ 3,500 g; >3,500 g ≤ 4,000 g; and >4,000 g). Additionally, cohort type (BirThree Cohort Study, CommCohort Study Type-1, and CommCohort Study Type-2) was added to the variables for clustering.

The missing variables used for clustering were imputed using the k-nearest neighbor (KNN) algorithm [36]. KNN selects k samples close to the missing values in the feature space and inputs the median of the k samples in the case of continuous variables or the most frequent category among the k samples in the case of categorical variables. KNN was implemented using the 'VIM' package in the R software (version 4.1.0) [37]. Based on previous reports [36,38], we set k to 219 as an odd number close to the square root of 48,365 participants.

2.3.2. Body mass index

BMI was computed by dividing the weight (kg) by the squared height (m^2) using self-reported height and weight on a questionnaire completed by the participants at baseline for each cohort. A BMI >25 kg/m^2 was defined as obesity based on the Western Pacific Region of the World Health Organization criteria for Asians [39].

2.3.3. Cluster analysis

The k-prototype is a clustering algorithm that combines k-means and k-modes and enables clustering using continuous and categorical variables [40]. The k-prototype was implemented using the 'clustMixType' package in R [41]. The number of clusters was set to five. Continuous variables were standardized by subtracting the mean of each variable and dividing it by the SD before clustering.

2.3.4. Genome-wide association studies

The GWASs with BMI as a continuous variable were conducted in two steps. Step-1: GWAS was conducted on all the 48,365 participants. Step-2:13,067 of the 48,365 participants were separated into five clusters using the k prototype. Thereafter, participants in each of the five obesity clusters and those with BMI <25 kg/m^2 were combined. A GWAS was conducted for each cluster (Fig. 2). To identify associations between autosomal SNPs and BMI, fastGWA with the GCTA software was employed [42]. FastGWA is a linear mixed model using a sparse genetic relationship matrix that is reportedly robust for population stratification and familial relationships [43]. The top 20 principal components computed from the PCA of the direct genotyping dataset, sex, age, and cohort type (BirThree Cohort Study, CommCohort Study Type-1, and CommCohort Study Type-2) were included as covariates. We set the Bonferroni genome-wide significance threshold at $P < 8.33 \times 10^{-9}$ ($5.0 \times 10^{-8}/6$), as six GWASs were conducted for Step-1 and -2. The detected SNPs were annotated using ANNOVAR [44]. Manhattan and quantile-quantile plots were generated using R software.

3. Results

3.1. Clustering

Following the assignment of 13,067 participants with obesity into five clusters, clusters 1–5 contained 628, 3,073, 4,111, 2,468, and 2,787 participants, respectively. Table S1 depicts the characteristics of the participants with obesity in each cluster. The variables were characterized by mean and standard deviation (SD) for continuous variables and number and percentage for categorical variables. The participants in Cluster 1 had the highest energy and nutrient intake and a higher frequency of leisure-time exercise. The participants in Cluster 2 were characterized by a higher proportion of older women and the highest percentage of non-smokers. The participants in Cluster 3 had the lowest energy and nutrient intake. A high proportion of participants did not exercise during leisure time or perform their usual physical activities (strenuous work, walking, or standing). The participants in Cluster 4 had the second-highest energy and nutrient intake. The participants in Cluster 5 were characterized by the largest proportion of men, the youngest age, and the longest time spent standing or sitting, and had the highest proportion of smokers, the highest proportion of alcohol drinkers, and the highest score on the psychological distress cale.

3.2. Gene interpretation

We observed several genes that satisfied the $P < 8.33 \times 10^{-9}$ threshold in Step-1 (Fig. 3 and Table S2). Most genes for which associations were detected in Step-1 were reportedly associated with obesity. Specifically, *LINC01741* [45–47] (Chromosome [Chr] 1), *CRYZL2P-SEC16B* [46,48] (Chr 1), *SEC16B* [47,47] (Chr 1), *TMEM18* [49,50] (Chr 2), *BDNF* [46,48] (Chr 11), *LINC00678* [51,52] (Chr 11), *BDNF-AS* [46,48] (Chr 11), *FTO* [9,10,46,53] (Chr 16), *MC4R* [54,55] (Chr 18), *GIPR* [14] (Chr 19), and *FBXO46* [51] (Chr 19) have been previously associated with BMI. Moreover, *KIF18A* (Chr 11) has been previously associated with visceral fat [56], *PMAIP1* (Chr 18) with serum IgE measurement [57] and monocyte count [58,59], *RSPH6A* (Chr 19) with high- and low-density lipoprotein cholesterol levels [60], *SYMPK* (Chr 19) with type 2 diabetes mellitus [61] and total cholesterol levels [51], and *FOXA3* (Chr 19) with waist-to-hip ratio adjusted for BMI [53].

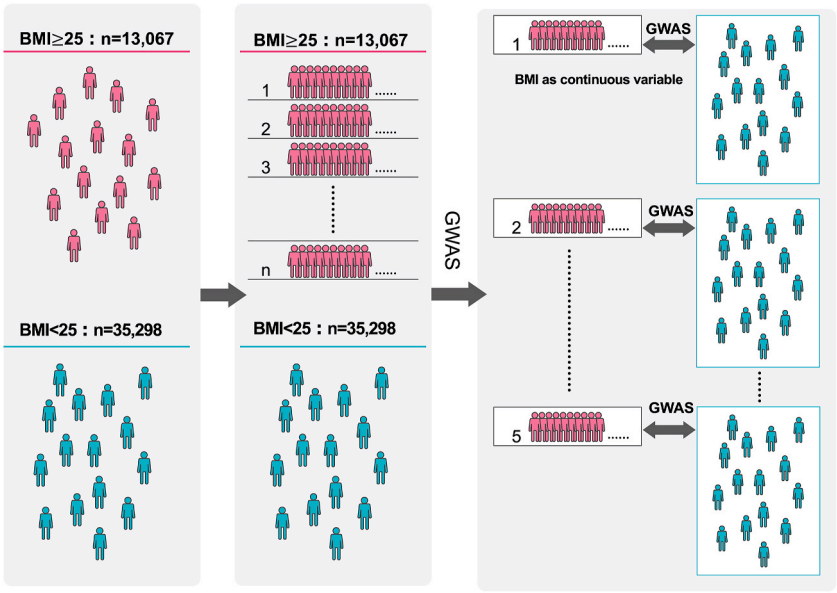

Fig. 2. Details of the cluster based GWAS.

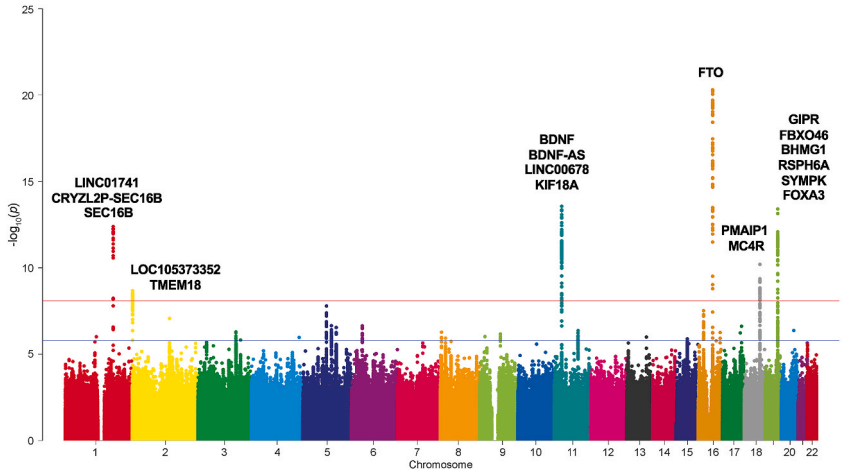

Fig. 3. Manhattan plot of Step-1 A genome-wide association study (GWAS) using body mass index (BMI) as a continuous variable is conducted on 48,365 participants.

Based on the GWAS results in Step-2, several variants detected in Step-1 were observed in separate clusters (Fig. 4 and Table S3). Genome-wide associations were not detected in Cluster 1. In Cluster 2, the loci that satisfied this threshold were *LINC01741*, *CRYZL2P-SEC16B* (Chr 1; intergenic), *CRYZL2P-SEC16B,* and *SEC16B* (Chr 1). In Cluster 3, *FTO* (Chr 16), *PMAIP1,* and *MC4R* (Chr 18; intergenic) loci were identified. In Cluster 4, *BDNF* (Chr 11), *BDNF-AS* (Chr 11), *BDNF-AS, LINC00678* (Chr 11), and *BDNF, KIF18A* (Chr 11) loci were identified. Additionally, in Cluster 5, *BDNF-AS, LINC00678* (Chr 11), *LINC00678* (Chr 11), *BDNF-AS* (Chr 11), *BDNF* (Chr 11), and *KIF18A* (Chr 11) loci were identified (Fig. 4). Quantile-quantile plots corresponding to the GWAS results of the main study are depicted in Fig. S3.

In the sub-analysis, the UK Biobank (UKB) data were used for comparison with the main study results. In Step-1, we verified the association between the representative obesity-related genes and BMI (Table S4 and Fig. S4). In Step-2, the clustering results for the 32,779 obese participants revealed that Clusters 1–5 comprised 5,874, 6,497, 6,919, 6,733, and 6756 participants, respectively. The characteristics of each cluster are presented in Table S5. In the GWAS results for Step-2, several variants detected in Step-1 were found in separate clusters. This was similar to the results of the Tohoku Medical Megabank Project (TMM) cohort analysis (Table S6 and Fig. S5).

4. Discussion

Herein, we conducted a GWAS on the participants according to their BMI in Step-1. In Step-2, participants with obesity were divided into five clusters based on obesity-related factors. A GWAS was conducted for each cluster. Consequently, several genes identified in previous studies were verified in Step-1. Of the 18 genes detected in Step-1, *LINC01741, CRYZL2P-SEC16B*, and *SEC16B* were significantly associated with Cluster 2. *FTO, PMAIP1*, and *MC4R* were linked to Cluster 3. *BDNF, BDNF-AS, LINC00678*, and *KIF18A* were associated with Clusters 4 and 5. A similar phenomenon was observed in the sub-analysis using UKB data, wherein unique obesity-related genes were detected in each cluster.

It is important to consider how the cluster characteristics relate to the variants identified in each cluster. The GWAS results in Step-2 may be partially elucidated by cluster characteristics. In Cluster 1, significant associations were not detected. This might be due to the low number of participants with BMI $>$25.0 kg/m^2 as this cluster contained the fewest participants with obesity. Hence, the detection power was insufficient.

FTO, *PMAIP1,* and *MC4R* (intergenic) variants were associated with BMI in Cluster 3. Variants in the *FTO* region regulate *IRX3* and *IRX5* expression [62]. These promote fat accumulation and cause obesity. Moreover, melanocortin-4-receptors (MC4R), transcribed by the *MR4C* gene, regulate food intake and energy expenditure [63,64]. MC4R in the paraventricular hypothalamus or amygdala controls food intake, whereas its expression elsewhere is responsible for energy expenditure [63]. Therefore, the genetic variants in *FTO* and *MC4R,* which have been reported to increase body fat accumulation and reduce energy expenditure, may partially account for the obesity of individuals in Cluster 3 despite low energy intake.

In Clusters 4 and 5, *BDNF* and *BDNF-AS* variants were identified. The participants with obesity in Cluster 4 had the second highest energy and nutrient intake, whereas those in Cluster 5 had the highest mean psychological distress score (K6 total score). Brain-derived neurotrophic factor (BDNF), which is transcribed by the *BDNF* gene, promotes the development and growth of nerve cells and has anti-obesity effects [65,66]. Furthermore, transcription of the *BDNF-AS* (antisense RNA) gene regulates *BDNF* expression [67]. Thus, altering BDNF regulation may affect the central nervous system and alter eating behaviors and psychiatric conditions, as observed in this cluster.

In Cluster 2, *SEC16B* variants were detected. Participants with obesity in this cluster had the highest proportion of women who were older and were nonsmokers. Variants of *SEC16B* may be associated with obesity via the regulation of dietary lipid absorption and

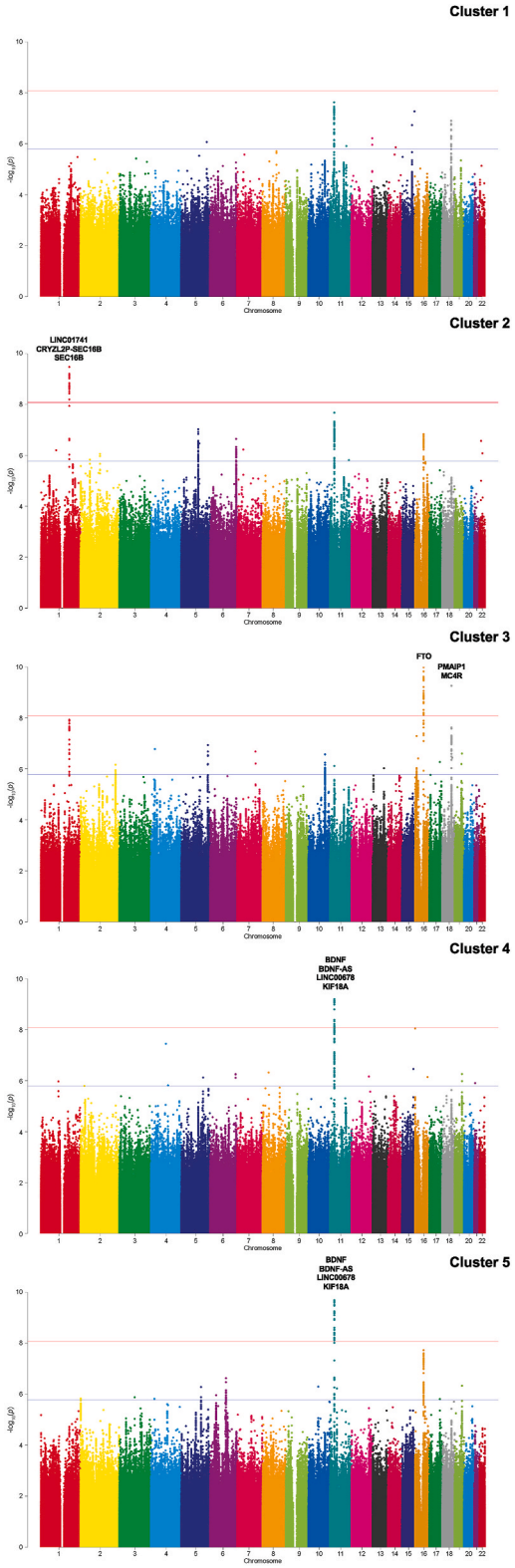

Fig. 4. Manhattan plots of Step-2 We have clustered 13,067 of 48,365 individuals with a body mass index (BMI) $\geq$ 25 kg/m^2 using the k-prototype. Subsequently, participants with obesity in each of the five clusters and those with a BMI <25 kg/m^2 are combined. A cluster-based genome-wide association study (cGWAS) performed according to the five clusters.

appetite [68,69]. To the best of our knowledge, no previous study has reported direct connections between the characteristics of Cluster 2 participants and *SEC16B* variants. However, it should be noted that the characteristics of clusters are not always recognizable by humans. Given that clustering algorithms extract latent features by combining numerous variables, the resulting clusters are not necessarily comprehensible, although they are more homogeneous. Thus, it is essential to define the obscure clusters identified using clustering algorithms.

This study has several strengths. The GWAS results had high validity. Most genes detected in this study were previously reported to be associated with BMI. Therefore, the GWAS data were considered appropriate. The TMM and UKB cohorts included diverse obesity-related factors. Using these two cohorts, it was possible to cluster the obese population into more homogeneous groups using a rich set of obesity-related factors. Sub-analysis replicated the phenomenon, wherein unique obesity-related genes were detected in each cluster. This finding supports the hypothesis that obesity comprises an aggregation of heterogeneous subgroups. Our findings suggest that dividing obese populations into homogeneous subpopulations would yield fewer genetic variants that could elucidate obesity in each subgroup. Although several issues remain to be addressed to elucidate the complete genetic architecture of obesity, this study provides important insights into the potential to inform the development of personalized treatments or nutritional support for obesity. More specifically, once clusters are identified, a classifier can be created using the cluster numbers as training data, which can then be applied to classify obesity into subgroups and verify the effectiveness of obesity treatment according to these subgroups.

This study had certain limitations. It is unclear whether the selection of variables, algorithms, or number of clusters is optimal. Herein, several obesity-related factors were selected. However, the existence of unknown obesity-related factors cannot be ruled out. Additionally, the number of clusters in this study was arbitrarily set to five, which should be explored in the future. Obesity was assessed at a temporal point; therefore, misclassification may have occurred. Even those who were not obese at the time of measurement have the potential to develop obesity with age. The BMI was computed using height and weight from self-reported questionnaires in the main study. Previous studies have demonstrated no substantial differences between BMI computed from self-reported or measured height and weight, indicating the usefulness of self-reported data [70]. Therefore, it is unlikely that the use of self-reported height and weight data significantly distorted our results. However, the impact of self-report bias in this study cannot be ruled out, as all data on obesity-related factors, not just weight and height, were collected through self-report questionnaires. Despite our best efforts to select similar variables for obesity-related factors in the TMM and UKB, we were unable to perfectly match obesity-related factors. This may have made it difficult to replicate the results of this study. Clustering the obese population reduces the sample size, leading to lower statistical power. Additionally, we could not assess the heritability of each cluster because of the small sample size. In the future, computing heritability in a population with a larger sample size and demonstrating that heritability is greater post clustering could provide strong evidence to indicate that obesity is an aggregation of heterogeneous genetic populations.

5. Conclusion

Our data suggest that a decreased sample size with increased homogeneity may provide insights into the genetic architecture of obesity.

Ethics declarations

This study was conducted in accordance with the guidelines of the Declaration of Helsinki. The protocol was reviewed and approved by the Institutional Review Board of the Tohoku Medical Megabank Organization (Approval number: 2022-4-089).

Funding statement

The Tohoku Medical Megabank Project Birth and Three-Generation Cohort Study and Community-Based Cohort Study were supported by the Japan Agency for Medical Research and Development (AMED) (grant numbers JP20km0105001 and JP21km0105002). This study was also supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), KAKENHI, Japan (grant numbers 19H03894 and 22H03346). AMED and MEXT played no role in the design or execution of the study. This research was supported in part by the Japan Agency for Medical Research and Development (AMED) under grant number JP21tm0424601.

Data statement

For the TMM Biobank, data are available from the authors upon reasonable request and with the permission of the TMM Biobank. All inquiries regarding access to data should be addressed to TMM Biobank at dist@megabank.tohoku.ac.jp. The data for the UKB are available to the public upon request from the UKB.

CRediT authorship contribution statement

Ippei Takahashi: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Hisashi Ohseto:** Writing – review & editing, Methodology, Conceptualization. **Fumihiko Ueno:** Writing – review & editing, Conceptualization. **Tomomi Oonuma:** Writing – review & editing, Conceptualization. **Akira Narita:** Writing – review & editing, Methodology, Conceptualization. **Taku Obara:** Writing – review & editing, Project administration, Investigation, Funding

acquisition, Data curation. **Mami Ishikuro:** Writing – review & editing. **Keiko Murakami:** Writing – review & editing. **Aoi Noda:** Writing – review & editing. **Atsushi Hozawa:** Writing – review & editing, Project administration, Investigation, Data curation. **Junichi Sugawara:** Writing – review & editing. **Gen Tamiya:** Writing – review & editing, Methodology, Conceptualization. **Shinichi Kuriyama:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:Shinichi Kuriyama reports financial support was provided by Japan Agency for Medical Research and Development (AMED) and Ministry of Education, Culture, Sports, Science and Technology (MEXT). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank the participants of the Tohoku Medical Megabank Project Birth and Three-Generation Cohort Study, Community-Based Cohort Study, and UK Biobank. The authors would also like to thank the staff members of the Tohoku Medical Megabank Organization (https://www.megabank.tohoku.ac.jp/english/a210901/) and UK Biobank.

Abbreviations

AS	Antisense RNA
BDNF	Brain-Derived Neurotrophic Factor
BirThree Cohort Study	TMM Birth and Three-Generation Cohort Study
BMI	Body Mass Index
cGWAS	Cluster-Based GWAS
Chr	Chromosome
CommCohort Study	TMM Community-Based Cohort Study
FFQ	Food Frequency Questionnaire
GRM	Genetic relationship matrix
GWAS	Genome-Wide Association Study
AMED	Japan Agency for Medical Research and Development
KNN	k-Nearest Neighbor
MAF	Minor allele frequency
MC4R	: MelanoCortin-4-Receptors
PCA	Principal Component Analysis
SD	Standard Deviation
SNPs	Single Nucleotide Polymorphisms
TMM	Tohoku Medical Megabank Project
UKB	UK Biobank

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e36023.

References

[1] D.W. Haslam, W.P. James, Obesity, Lancet 366 (2005) 1197–1209, https://doi.org/10.1016/S0140-6736(05)67483-1.
[2] R.H. Eckel, S.M. Grundy, P.Z. Zimmet, The metabolic syndrome, Lancet 365 (2005) 1415–1428, https://doi.org/10.1016/S0140-6736(05)66378-7.
[3] M. Ng, T. Fleming, M. Robinson, B. Thomson, N. Graetz, C. Margono, E.C. Mullany, S. Biryukov, C. Abbafati, S.F. Abera, et al., Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013, Lancet 384 (2014) 766–781, https://doi.org/10.1016/S0140-6736(14)60460-8.
[4] X. Lin, H. Li, Obesity: epidemiology, pathophysiology, and therapeutics, Front. Endocrinol. 12 (2021) 706978, https://doi.org/10.3389/fendo.2021.706978.
[5] H.N. Lyon, J.N. Hirschhorn, Genetics of common forms of obesity: a brief overview, Am. J. Clin. Nutr. 82 (Suppl) (2005) 215S–217S, https://doi.org/10.1093/ajcn/82.1.215S.
[6] R. Feng, How much do we know about the heritability of BMI? Am. J. Clin. Nutr. 104 (2016) 243–244, https://doi.org/10.3945/ajcn.116.139451.
[7] C.E. Elks, M. den Hoed, J.H. Zhao, S.J. Sharp, N.J. Wareham, R.J. Loos, K.K. Ong, Variability in the heritability of body mass index: a systematic review and meta-regression, Front. Endocrinol. 3 (2012) 29, https://doi.org/10.3389/fendo.2012.00029.
[8] J. Min, D.T. Chiu, Y. Wang, Variation in the heritability of body mass index based on diverse twin studies: a systematic review, Obes. Rev. 14 (2013) 871–882, https://doi.org/10.1111/obr.12065.
[9] M. Akiyama, Y. Okada, M. Kanai, A. Takahashi, Y. Momozawa, M. Ikeda, N. Iwata, S. Ikegawa, M. Hirata, K. Matsuda, et al., Genome-wide association study identifies 112 new loci for body mass index in the Japanese population, Nat. Genet. 49 (2017) 1458–1467, https://doi.org/10.1038/ng.3951.

[10] A.E. Locke, B. Kahali, S.I. Berndt, A.E. Justice, T.H. Pers, F.R. Day, C. Powell, S. Vedantam, M.L. Buchkovich, J. Yang, et al., Genetic studies of body mass index yield new insights for obesity biology, Nature 518 (2015) 197–206, https://doi.org/10.1038/nature14177.
[11] L. Yengo, J. Sidorenko, K.E. Kemper, Z. Zheng, A.R. Wood, M.N. Weedon, T.M. Frayling, J. Hirschhorn, J. Yang, P.M. Visscher, GIANT Consortium, Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry, Hum. Mol. Genet. 27 (2018) 3641–3649, https://doi.org/10.1093/hmg/ddy271.
[12] J. Yang, A. Bakshi, Z. Zhu, G. Hemani, A.A. Vinkhuyzen, S.H. Lee, M.R. Robinson, J.R. Perry, I.M. Nolte, J.V. van Vliet-Ostaptchouk, et al., Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index, Nat. Genet. 47 (2015) 1114–1120, https://doi.org/10.1038/ng.3390.
[13] E.K. Speliotes, C.J. Willer, S.I. Berndt, K.L. Monda, G. Thorleifsson, A.U. Jackson, H. Lango Allen, C.M. Lindgren, J. Luan, R. Mägi, et al., Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index, Nat. Genet. 42 (2010) 937–948, https://doi.org/10.1038/ng.686.
[14] W. Wen, W. Zheng, Y. Okada, F. Takeuchi, Y. Tabara, J.Y. Hwang, R. Dorajoo, H. Li, F.J. Tsai, X. Yang, et al., Meta-analysis of genome-wide association studies in East Asian-ancestry populations identifies four new loci for body mass index, Hum. Mol. Genet. 23 (2014) 5492–5504, https://doi.org/10.1093/hmg/ddu248.
[15] A. Scuteri, S. Sanna, W.M. Chen, M. Uda, G. Albai, J. Strait, S. Najjar, R. Nagaraja, M. Orrú, G. Usala, et al., Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits, PLoS Genet. 3 (2007) e115, https://doi.org/10.1371/journal.pgen.0030115.
[16] A.V. Khera, M. Chaffin, K.H. Wade, S. Zahid, J. Brancale, R. Xia, M. Distefano, O. Senol-Cosar, M.E. Haas, et al., Polygenic prediction of weight and obesity trajectories from birth to adulthood, Cell 177 (2019) 587–596.e9, https://doi.org/10.1016/j.cell.2019.03.028.
[17] M. Traylor, H. Markus, C.M. Lewis, Homogeneous case subgroups increase power in genetic association studies, Eur. J. Hum. Genet. 23 (2015) 863–869, https://doi.org/10.1038/ejhg.2014.194.
[18] F. Ueno, T. Onuma, I. Takahashi, H. Ohseto, A. Narita, T. Obara, M. Ishikuro, K. Murakami, A. Noda, F. Matsuzaki, et al., Deep embedded clustering by relevant scales and genome-wide association study in autism, bioRxiv (2022), https://doi.org/10.1101/2022.07.25.500917. Published online July 25.
[19] A. Narita, M. Nagai, S. Mizuno, S. Ogishima, G. Tamiya, M. Ueki, R. Sakurai, S. Makino, T. Obara, M. Ishikuro, et al., Clustering by phenotype and genome-wide association study in autism, Transl. Psychiatry 10 (2020) 290, https://doi.org/10.1038/s41398-020-00951-x.
[20] World Medical Association, World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects, JAMA 310 (2013) 2191–2194, https://doi.org/10.1001/jama.2013.281053.
[21] S. Kuriyama, N. Yaegashi, F. Nagami, T. Arai, Y. Kawaguchi, N. Osumi, M. Sakaida, Y. Suzuki, K. Nakayama, H. Hashizume, et al., The tohoku medical megabank Project: design and mission, J. Epidemiol. 26 (2016) 493–511, https://doi.org/10.2188/jea.JE20150268.
[22] S. Kuriyama, H. Metoki, M. Kikuya, T. Obara, M. Ishikuro, C. Yamanaka, M. Nagai, H. Matsubara, T. Kobayashi, et al., Cohort profile: tohoku medical megabank Project birth and three-generation cohort study (TMM BirThree cohort study): rationale, progress and perspective, Int. J. Epidemiol. 49 (2020) 18–19m, https://doi.org/10.1093/ije/dyz169.
[23] A. Hozawa, K. Tanno, N. Nakaya, T. Nakamura, N. Tsuchiya, T. Hirata, A. Narita, M. Kogure, K. Nochioka, R. Sasaki, et al., Study profile of the Tohoku medical Megabank community-based cohort study, J. Epidemiol. 31 (2021) 65–76, https://doi.org/10.2188/jea.JE20190271.
[24] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al., UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age, PLoS Med. 12 (2015) e1001779, https://doi.org/10.1371/journal.pmed.1001779.
[25] C. Bycroft, C. Freeman, D. Petkova, G. Band, L.T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J.O. Connell, et al., Genome-wide genetic data on ~500,000 UK Biobank participants, bioRxiv (2017), https://doi.org/10.1101/166298. Published online July 20.
[26] C. Bycroft, C. Freeman, D. Petkova, G. Band, L.T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, et al., The UK biobank resource with deep phenotyping and genomic data, Nature 562 (2018) 203–209, https://doi.org/10.1038/s41586-018-0579-z.
[27] M. Yamada, I.N. Motoike, K. Kojima, N. Fuse, A. Hozawa, S. Kuriyama, F. Katsuoka, S. Tadaka, M. Shirota, M. Sakurai, et al., Genetic loci for lung function in Japanese adults with adjustment for exhaled nitric oxide levels as airway inflammation indicator, Commun. Biol. 4 (2021) 1288, https://doi.org/10.1038/s42003-021-02813-8.
[28] C.C. Chang, C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, J.J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets, GigaScience 4 (2015) 7, https://doi.org/10.1186/s13742-015-0047-8.
[29] O. Delaneau, J.F. Zagury, J. Marchini, Improved whole-chromosome phasing for disease and population genetic studies, Nat. Methods 10 (2013) 5–6, https://doi.org/10.1038/nmeth.2307.
[30] J. O'Connell, D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, M. Cocca, M. Traglia, J. Huang, J.E. Huffman, I. Rudan, et al., A general approach for haplotype phasing across the full spectrum of relatedness, PLoS Genet. 10 (2014) e1004234, https://doi.org/10.1371/journal.pgen.1004234.
[31] S. Tadaka, F. Katsuoka, M. Ueki, K. Kojima, S. Makino, S. Saito, A. Otsuki, C. Gocho, M. Sakurai-Yageta, et al., 3.5KJPNv2: an allele frequency panel of 3552 Japanese individuals including the X chromosome, Hum. Genome Var. 6 (2019) 28, https://doi.org/10.1038/s41439-019-0059-5.
[32] 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, et al., A global reference for human genetic variation, Nature 526 (2015) 68–74, https://doi.org/10.1038/nature15393.
[33] B.N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, PLoS Genet. 5 (2009) e1000529.
[34] R.C. Kessler, G. Andrews, L.J. Colpe, E. Hiripi, D.K. Mroczek, S.L. Normand, E.E. Walters, A.M. Zaslavsky, Short screening scales to monitor population prevalences and trends in non-specific psychological distress, Psychol. Med. 32 (2002) 959–976, https://doi.org/10.1017/s0033291702006074.
[35] T.A. Furukawa, N. Kawakami, M. Saitoh, Y. Ono, Y. Nakane, Y. Nakamura, H. Tachimori, N. Iwata, H. Uda, H. Nakane, et al., The performance of the Japanese version of the K6 and K10 in the World mental health survey Japan. Int. J. Methods psychiatr, Res. 17 (2008) 152–158, https://doi.org/10.1002/mpr.257.
[36] Z. Zhang, Introduction to machine learning: k-nearest neighbors, Ann. Transl. Med. 4 (2016) 218, https://doi.org/10.21037/atm.2016.03.37.
[37] M. Templ, A. Kowarik, A. Alfons, G. de Cilia, B. Prantner, W. Rannetbauer, Visualization and imputation of missing values. https://cran.r-project.org/web/packages/VIM, 2022. (Accessed 8 July 2024).
[38] B. Lantz, Machine Learning with R: Expert Techniques for Predictive Modeling to Solve All Your Data Analysis Problems, second ed., Packt Publishing, 2015. https://www.packtpub.com/en-us/product/machine-learning-with-r-9781784393908. (Accessed 8 July 2024).
[39] World Health Organization. Regional office for the western pacific. The asia-pacific perspective: redefining obesity and its treatment, Sydney: Health Communications, 2000. https://iris.who.int/handle/10665/206936 (Accessed 8 July 2024).
[40] Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values, Data Min. Knowl. Discov. 2 (1998) 283–304, https://doi.org/10.1023/A:1009769707641.
[41] Package 'clustMixType' k-prototypes clustering for mixed variable-type data. https://cran.r-project.org/web/packages/clustMixType/clustMixType.pdf, 2021. (Accessed 8 July 2024).
[42] Y. Wang, X. Ding, Z. Tan, C. Ning, K. Xing, T. Yang, Y. Pan, D. Sun, C. Wang, Genome-wide association study of piglet uniformity and farrowing interval, Front. Genet. 8 (2017) 194, https://doi.org/10.3389/fgene.2017.00194.
[43] L. Jiang, Z. Zheng, H. Fang, J. Yang, A generalized linear mixed model association tool for biobank-scale data, Nat. Genet. 53 (2021) 1616–1621, https://doi.org/10.1038/s41588-021-00954-4.
[44] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res. 38 (2010) e164, https://doi.org/10.1093/nar/gkq603.
[45] M.C.Y. Ng, M. Graff, Y. Lu, A.E. Justice, P. Mudgal, C.T. Liu, K. Young, L.R. Yanek, M.F. Feitosa, M.K. Wojczynski, et al., Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of African ancestry: african Ancestry Anthropometry Genetics Consortium, PLoS Genet. 13 (2017) e1006719, https://doi.org/10.1371/journal.pgen.1006719.

[46] G.L. Wojcik, M. Graff, K.K. Nishimura, R. Tao, J. Haessler, C.R. Gignoux, H.M. Highland, Y.M. Patel, E.P. Sorokin, C.L. Avery, et al., Genetic analyses of diverse populations improves discovery for complex traits, Nature 570 (2019) 514–518, https://doi.org/10.1038/s41586-019-1310-4.
[47] K.L. Monda, G.K. Chen, K.C. Taylor, C. Palmer, T.L. Edwards, L.A. Lange, M.C. Ng, A.A. Adeyemo, M.A. Allison, L.F. Bielak, et al., A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry, Nat. Genet. 45 (2013) 690–696, https://doi.org/10.1038/ng.2608.
[48] G. Thorleifsson, G.B. Walters, D.F. Gudbjartsson, V. Steinthorsdottir, P. Sulem, A. Helgadottir, U. Styrkarsdottir, S. Gretarsdottir, S. Thorlacius, I. Jonsdottir, et al., Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity, Nat. Genet. 41 (2009) 18–24, https://doi.org/10.1038/ng.274.
[49] Y.F. Pei, L. Zhang, Y. Liu, J. Li, H. Shen, Y.Z. Liu, Q. Tian, H. He, S. Wu, S. Ran, et al., Meta-analysis of genome-wide association data identifies novel susceptibility loci for obesity, Hum. Mol. Genet. 23 (2014) 820–830, https://doi.org/10.1093/hmg/ddt464.
[50] C.J. Willer, E.K. Speliotes, R.J. Loos, S. Li, C.M. Lindgren, I.M. Heid, S.I. Berndt, A.L. Elliott, A.U. Jackson, C. Lamina, et al., Six new loci associated with body mass index highlight a neuronal influence on body weight regulation, Nat. Genet. 41 (2009) 25–34, https://doi.org/10.1038/ng.287.
[51] Z. Zhu, Y. Guo, H. Shi, C.L. Liu, R.A. Panganiban, W. Chung, L.J. O'Connor, B.E. Himes, S. Gazal, K. Hasegawa, et al., Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank, J. Allergy Clin. Immunol. 145 (2020) 537–549, https://doi.org/10.1016/j.jaci.2019.09.035.
[52] I. Tachmazidou, D. Süveges, J.L. Min, G.R.S. Ritchie, J. Steinberg, K. Walter, V. Iotchkova, J. Schwartzentruber, J. Huang, Y. Memari, et al., Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits, Am. J. Hum. Genet. 100 (2017) 865–884, https://doi.org/10.1016/j.ajhg.2017.04.014.
[53] S. Sakaue, M. Kanai, Y. Tanigawa, J. Karjalainen, M. Kurki, S. Koshiba, A. Narita, T. Konuma, K. Yamamoto, M. Akiyama, et al., A cross-population atlas of genetic associations for 220 human phenotypes, Nat. Genet. 53 (2021) 1415–1424, https://doi.org/10.1038/s41588-021-00931-x.
[54] A.R. Barton, M.A. Sherman, R.E. Mukamel, P.R. Loh, Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses, Nat. Genet. 53 (2021) 1260–1269, https://doi.org/10.1038/s41588-021-00892-1.
[55] P. Akbari, A. Gilani, O. Sosina, J.A. Kosmicki, L. Khrimian, Y.Y. Fang, T. Persaud, V. Garcia, D. Sun, A. Li, et al., Sequencing of 640,000 exomes identifies GPR75 variants associated with protection from obesity, Science 373 (2021) eabf8683.
[56] J. Shin, C. Syme, D. Wang, L. Richer, G.B. Pike, D. Gaudet, T. Paus, Z. Pausova, Novel genetic locus of visceral fat and systemic inflammation, J. Clin. Endocrinol. Metab. 104 (2019) 3735–3742, https://doi.org/10.1210/jc.2018-02656.
[57] A.T. Akenroye, T. Brunetti, K. Romero, M. Daya, K. Kanchan, G. Shankar, S. Chavan, M. Preethi Boorgula, E.A. Ampleford, H.F. Fonseca, et al., Genome-wide association study of asthma, total IgE, and lung function in a cohort of Peruvian children, J. Allergy Clin. Immunol. 148 (2021) 1493–1504, https://doi.org/10.1016/j.jaci.2021.02.035.
[58] M.H. Chen, L.M. Raffield, A. Mousas, S. Sakaue, J.E. Huffman, A. Moscati, B. Trivedi, T. Jiang, P. Akbari, D. Vuckovic, et al., Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations, Cell 182 (2020) 1198–1213.e14, https://doi.org/10.1016/j.cell.2020.06.045.
[59] D. Vuckovic, E.L. Bao, P. Akbari, C.A. Lareau, A. Mousas, T. Jiang, M.H. Chen, L.M. Raffield, M. Tardaguila, J.E. Huffman, et al., The polygenic and monogenic basis of blood traits and diseases, Cell 182 (2020) 1214–1231.e11, https://doi.org/10.1016/j.cell.2020.08.008.
[60] N. Sinnott-Armstrong, Y. Tanigawa, D. Amar, N. Mars, C. Benner, M. Aguirre, G.R. Venkataraman, M. Wainberg, H.M. Ollila, T. Kiiskinen, et al., Genetics of 35 blood and urine biomarkers in the UK Biobank, Nat. Genet. 53 (2021) 185–194.
[61] A. Mahajan, D. Taliun, M. Thurner, N.R. Robertson, J.M. Torres, N.W. Rayner, A.J. Payne, V. Steinthorsdottir, R.A. Scott, N. Grarup, et al., Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps, Nat. Genet. 50 (2018) 1505–1513.
[62] M. Claussnitzer, S.N. Dankel, K.H. Kim, G. Quon, W. Meuleman, C. Haugen, V. Glunk, I.S. Sousa, J.L. Beaudry, V. Puviindran, et al., FTO obesity variant circuitry and adipocyte browning in humans, N. Engl. J. Med. 373 (2015) 895–907, https://doi.org/10.1056/NEJMoa1502214.
[63] N. Balthasar, L.T. Dalgaard, C.E. Lee, J. Yu, H. Funahashi, T. Williams, M. Ferreira, V. Tang, R.A. McGovern, C.D. Kenny, et al., Divergence of melanocortin pathways in the control of food intake and energy expenditure, Cell 123 (2005) 493–505, https://doi.org/10.1016/j.cell.2005.08.035.
[64] M.J. Krashes, B.B. Lowell, A.S. Garfield, Melanocortin-4 receptor–regulated energy homeostasis, Nat. Neurosci. 19 (2016) 206–219, https://doi.org/10.1038/nn.4202.
[65] E.E. Noble, C.J. Billington, C.M. Kotz, C. Wang, The lighter side of BDNF, Am. J. Physiol. Regul. Integr. Comp. Physiol. 300 (2011) R1053–R1069, https://doi.org/10.1152/ajpregu.00776.2010.
[66] M. Pandit, T. Behl, M. Sachdeva, S. Arora, Role of brain derived neurotropic factor in obesity, Obes. Med. 17 (2020) 100189, https://doi.org/10.1016/j.obmed.2020.100189.
[67] S. Ghafouri-Fard, T. Khoshbakht, M. Taheri, M. Ghanbari, A concise review on the role of BDNF-AS in human disorders, Biomed. Pharmacother. 142 (2021) 112051, https://doi.org/10.1016/j.biopha.2021.112051.
[68] R. Shi, W. Lu, Y. Tian, B. Wang, L. Ave, Intestinal SEC16B modulates obesity by controlling dietary lipid absorption, bioRxiv (2021), https://doi.org/10.1101/2021.12.07.471468. Published online Dec 7.
[69] K. Hotta, M. Nakamura, T. Nakamura, T. Matsuo, Y. Nakata, S. Kamohara, N. Miyatake, K. Kotani, R. Komatsu, N. Itoh, et al., Association between obesity and polymorphisms in SEC16B, TMEM18, GNPDA2, BDNF, FAIM2 and MC4R in a Japanese population. J, Hum. Genet. 54 (2009) 727–731, https://doi.org/10.1038/jhg.2009.106.
[70] L.A.H. Haakstad, T. Stensrud, C. Gjestvang, Does self-perception equal the truth when judging own body weight and height? Int. J. Environ. Res. Public Health 18 (2021) 8502, https://doi.org/10.3390/ijerph18168502.