**Anthropometric genetic risk scores replication analysis plan**

**Background**

We have conducted GWAS in China Kadoorie (CKB) and UK Biobank (UKB), followed by trans-ethnic (TE) meta-analysis, for anthropometric traits including:

* Body mass index (BMI): weight (kg) / height (m)2
* Waist circumference (WC): cm
* Hip circumference (HIP): cm
* Waist-hip ratio (WHR): WC / HIP

We have derived and tested genetic risk scores based on the variants identified from each of the CKB, UKB, or TE analyses for their predictive performance in CKB and UKB, including in non-European UKB populations. We have additionally derived two scores that are subsets of the UKB and TE GRSs that aim to explore the impact of the TE meta-analysis on the refinement of variant selection. In total we have scores for each trait composed of variants from:

* CKB GWAS results (CKB\_GRS)
* UKB GWAS results (UKB\_GRS)
* TE meta-analysis (TE\_GRS)
* Subset of UKB GWAS results matched to loci discovered in TE analysis (UKB\_finemap)
* Subset of TE results matched to loci discovered in UKB analysis (TE\_finemap)

We wish to replicate our findings on the improvements gained in trait prediction from TE meta-analysis in external datasets.

Please could you conduct the following analyses for whichever of the 4 traits listed above that you have available in your dataset.

**GRS construction**

We provide lists of the different scores we would like you to construct. For each trait we provide separate folders containing a .csv file for each risk score we would like you to assess. For example, the file named “BMI\_UKB\_GRS.csv” in the “BMI” folder lists all the variants that should be combined into a score for BMI based on variants identified in the UKB GWAS.

The files list SNP ID in the form CHR:BP:A1:A2 with the alleles ordered alphabetically (e.g. 1:33877057:G:T), CHR, BP(GRCh37), EA (effect allele), NEA (non-effect allele), and the variant’s weighting. Please ensure the variants in your dataset have the same alleles as those listed. We have not provided rsids for the variants as there may be mismatches between studies in the rsid assigned to any given variant.

If a variant is not present in your dataset (e.g. was not imputed) please simply exclude it from all analyses.

Variants listed in the “finemap\_GRS.csv” files are paired based on locus identifier - indicated under column 'loc'. For example, SNP 1:19934900:A:G listed in *BMI\_UKB\_finemap\_GRS.csv* and SNP 1:19922306:A:AT listed in *BMI\_TE\_finemap\_GRS.csv* both have the value “BMI.chr.1.10” in the ‘loc’ column, indicating they lie within the same locus.

When compiling the finemap scores for each trait, if one of the listed variants is not present in your dataset please exclude all variants from the scores that have the same ‘loc’ value in both the TE and UKB finemap files. For example, if 2 variants are listed as having the same loc value in the *BMI\_UKB\_finemap\_GRS.csv* file and one of these variants is not present in your data, please exclude the second variant that is present from the BMI\_UKB\_finemap score and also exclude the 2 variants that have the same loc value listed in the *BMI\_TE\_finemap\_GRS.csv* from the BMI\_TE\_finemap score.

When constructing each score please code your dosage data to be on the effect alleles listed in the relevant GRS file, and construct the scores in sexes combined using the corresponding weights as listed in the GRS files. Scores should be derived in the usual manner, as the weighted sum of each individual’s effect allele dosages, i.e.

sum(weight \* effect allele count)

For imputed variants, please use (non-integer) effect allele counts derived as the sum of the imputed genotype probabilities

**Partial r2 estimation**

Please derive the estimated variance explained by each GRS, as a partial r2 between two models with and without the GRS.

For each model, please use unrelated individuals, excluded using e.g. a KING cutoff of 0.05 for identity by descent as implemented in PLINK2, or as appropriate for your study.

For each trait separately:

* Fit a linear model regressing the trait against age, age2, sex, and the top 10 study principal components (PCs) or as appropriate for your dataset, plus any other appropriate study-specific covariates, i.e.

Model1 = BMI ~ age + age2 + sex + PC1 + PC2+ … + study-specific covariates

* Fit a second model that is the same as above but includes one of the corresponding GRSs for that trait eg:

Model2= BMI ~ BMI\_UKB\_GRS + age + age2 + sex + PC1 + PC2+ … + study-specific covariates

The partial r2 can then be estimated as the ratio of the difference between the error sum of squares of the reduced model and the full model to the error sum of squares of the reduced model:

(SSE (model1) - SSE (model2)) / SSE (model1)

This can be estimated in R using the rsq.partial() function from the rsq package:

Partial\_r2 <- rsq.partial(model2, model1)$partial.rsq

**Summary statistics**

For the partial r2 analysis, please fill in the file “GRS\_results\_table.csv” we have provided with the results for the different traits/scores. Please add the study name and the date of the analysis as a prefix to the file name, e.g. CKB\_01\_04\_21\_ GRS\_results\_table.csv. The results file has the columns:

* Cohort – Name of cohort
* Pop – Continental ancestry of the population, please use 1000 genomes abbreviations
* Trait – Trait assessed
* Score – Name of GRS file e.g. BMI\_UKB\_GRS
* R2 – Partial r2 value
* N – Sample size included in Model2
* N\_PC - Number of PCs included
* N\_snp – Number of variants listed in the GRS .csv file included in the GRS in your study
* Cov - Additional study specific covariates, NA if none were added

Please also complete the file “study\_descrip.csv” detailing the characteristics of your study for all traits present in your dataset stratified by sex. Please add the study name and the date of the analysis as a prefix to the file name, e.g. CKB\_01\_04\_21\_study\_descrip.csv. The file contains the columns:

* Cohort – Cohort name
* Sex – Sex of individuals
* N – Number of individuals in study population
* Age – Mean age in study population
* Age\_SD – Standard deviation of age in study population
* BMI – Mean BMI in study population
* BMI\_SD – Standard deviation of BMI in study population
* WC – Mean WC (cm) in study population
* WC\_SD – Standard deviation of WC in study population
* HIP – Mean HIP (cm) in study population
* HIP\_SD – Standard deviation of HIP in study population
* WHR – Mean WHR in study population
* WHR\_SD – Standard deviation of WHR in study population

Please also provide a single file documenting the SNP IDs (CHR:BP:A1:A2) of variants we listed for inclusion in any of the scores that were not present in your dataset. Please save this file as “study\_date\_missing.csv”, e.g. CKB\_01\_04\_21\_missing.csv.