The American Journal of Human Genetics, Volume *101*

Supplemental Data

Penetrance of Polygenic Obesity Susceptibility Loci across the Body Mass Index Distribution

Arkan Abadi, Akram Alyass, Sebastien Robiou du Pont, Ben Bolker, Pardeep Singh, Viswanathan Mohan, Rafael Diaz, James C. Engert, Salim Yusuf, Hertzel C. Gerstein, Sonia S. Anand, and David Meyre

SUPPLEMENTAL NOTE

Analytical Description: Ordinary least squares (OLS) regression is the classic method to estimate mean effects of SNPs on a quantitative trait. OLS models are particularly useful when the assumptions of linearity, normality, and homoscedasticity are met, but otherwise, require proper corrections in order to allow unbiased parameter estimation and valid inference. These models are developed on the basis of true fixed effects and do not capture true variability in the effects of genetic risk factors in the presence of single and mixed gene-environment (G$\times$E) and gene-gene (G$\times$G) interactions. If such interactions are unadjusted, OLS models will produce estimates with limited reproducibility that depend on the context of the sample population and the degree of exposure to interacting variables (e.g. environmental exposure).[1] Reproducibility is a well-known problem in genetic epidemiology for complex phenotypes that involve interactions.[2] Alternatively, GWAS may use case/control designs to compare BMI categories, where binary logistic regression is used to estimate the fixed effects of SNPs on the probability of belonging to either of two factor levels (e.g. normal-weight vs. obesity subgroups). However, subgroup analysis not only reduces statistical power due to loss of sample size and uneven group levels, but also limits interpretation to pair-wise comparisons. In addition, logistic regression profiles pre-selected segments of the BMI distribution, which can be problematic to assign a priori.

Conditional quantile regression (CQR) is an alternative regression technique that permits the assessment of associations at the full scope of the outcome distribution by examining the effects of regressors at a series of quantiles of the outcome distribution without dividing the sample into subgroups.[3,4] CQR models the effects of a change in one unit of a predictor on the position of a given quantile of the outcome. It also utilizes the entire data set for parameter estimation, confidence interval construction and hypothesis testing regardless of the specified quantiles and does not suffer the statistical limitations of subgroup analysis. This regression framework has recently gained traction in clinical epidemiology to generate fetal, childhood and adolescent growth curves.[5-7] Recent reports have highlighted the potential applications of CQR in genetic epidemiology.[8-10] To our knowledge, CQR has not been applied to model the variability in effect size estimates along the sample outcome distribution in the presence of single and mixed G$\times$E and G$\times$G interactions.

Variations in effect size estimates due to unadjusted interactions can be modelled using CQR as a re-formulation of heteroscedastic OLS models.[3,11,12] Lets consider a sample of n independent and identical distributed (i.i.d) variables $Y_1, ..., Y_n$ with $cdf\ F_Y(y)$, where $y_1, ..., y_n$ are their respective observed values. Let's also assume they follow a linear relationship with an interaction term given as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 g_i + \beta_3 x_i g_i + \epsilon_i \quad (1)$$

where x_i corresponds to the unknown/unmeasured interacting variable, $X \sim F_X(\mu_x, \sigma_x^2)$; g_i is the observed genotype of the genetic variant G under Hardy-Weinberg equilibrium (HWE) with a population allele frequency, p, where $G \sim B(2, p)$; and ϵ_i is the random

error with $\epsilon \sim F_{\epsilon}(0, \sigma_{\epsilon})$. The coefficients β_0, β_1, β_2 and β_3 represent the intercept, the marginal effect of X, the marginal effect of G and the interaction effect of G and X, respectively. The conditional distribution of the response variable Y can be described as $F_{Y|X=x,G=g} \sim \big(\beta_0 + \beta_1 x + \beta_2 g + \beta_3 xg, \sigma_y^2\big)$. If the interacting variable, X, is not adjusted then the conditional density of Y given G can be shown to have a mean and variance:

$$E\left[Y \mid G = g\right] = (\beta_0 + \beta_1 \mu_x) + (\beta_2 + \beta_3 \mu_x) g \tag{2}$$

$$Var(Y \mid G = g) = \sigma_x^2 (\beta_1 + \beta_3 g)^2 + \sigma_{\epsilon}^2 \tag{3}$$

The resulting conditional distribution $F_{Y|G}(y \mid G = g)$ simply translates to a heteroscedastic linear model with partitioned residuals where $\sigma(g) = (\beta_1 + \beta_3 g)$. That is

$$y_i = \big(\beta_0 + \beta_1 \mu_x\big) + \big(\beta_2 + \beta_3 \mu_x\big) g_i + \big(\beta_1 + \beta_3 g_i\big) \epsilon_{i,1} + \epsilon_{i,2} \tag{4}$$

where $\epsilon_1 \sim F_{\epsilon_1}(0, \sigma_x^2)$ and $\epsilon_2 \sim F_{\epsilon_2}(0, \sigma_{\epsilon}^2)$. The conditional quantile function for the heteroscedastic model under i.i.d errors is

$$\begin{aligned} Q_Y(\tau \mid G = g) &= \big[\beta_0 + \beta_1 \mu_x + \beta_1 Q_{\epsilon_1}(\tau) + Q_{\epsilon_2}(\tau)\big] + g\big[\beta_2 + \beta_3 \mu_x + \beta_3 Q_{\epsilon_1}(\tau)\big] \\ &= \beta_0^*(\tau) + \beta_1^*(\tau) g \end{aligned} \tag{5}$$

which is a CQR model with the true fixed parameters $\beta_0^*(\tau)$ and $\beta_1^*(\tau)$. τ can be any quantile of the sample outcome distribution of Y. This formulation can be generalized further for a set of k independent interacting variables in matrix form as

$$Q_Y(\tau \mid G = g) = A\beta(\tau) \tag{6}$$

where $A \in \mathbb{R}^{n \times 2}$ is the design matrix $\big[1, G^{'}\big]$ and

$$\beta(\tau) = \begin{pmatrix} \beta_0 + \sum_{j=1}^{k} \beta_{x_j} \mu_j \\ \beta_g + \sum_{j=1}^{k} \beta_{int_j} \mu_j \end{pmatrix} + \begin{pmatrix} \sum_{j=1}^{k} \beta_{x_j} Q_{\epsilon_j}(\tau) + Q_{\epsilon_{k+1}}(\tau) \\ \sum_{j=1}^{k} \beta_{int_j} Q_{\epsilon_j}(\tau) \end{pmatrix} \tag{7}$$

Here, β_g and β_{x_j} are the main effects of the genetic variant and $j \in 1, .., j$ are the unknown/unmeasured variables with their respective interaction coefficients β_{int_j}. The cumulative two-way interactions of k variables results in a linear function with τ as a result of the symmetric heteroscedasticity function $^{'}\boldsymbol{\gamma}$ where $\in \mathbb{R}^{k \times 1}$ and γ has elements $\sigma_j(g) = \beta_{x_j} + \beta_{int_j} g$. Under an additive genetic model, the main effect of the genetic variant $\beta(\tau)$ is a fixed constant for all $\boldsymbol{\tau} \in \tau_1, ..., \tau_m$ if and only if all interacting effects are zero, i.e. $\beta_{int_j} = 0$. It is possible to further break down the independence assumption between interacting variables using a variance-covariance matrix of partial errors, but the above formulation serves as a simple analytical demonstration for the use of CQR in modelling unadjusted interactions. A linear trend of estimates with τ corresponds to cumulative two-way interactions, while quadratic curves supports complex higher order interactions. Hence, the association of genetic variants under unadjusted interacting variables simply reduces to the modelling of CQR estimates along the distribution of the outcome at $\boldsymbol{\tau} \in \tau_1, ..., \tau_m$.

This is accomplished by using meta-regression (MR) to model the heterogeneity of CQR estimates across the sample outcome distribution and estimate the change in CQR estimates with $\boldsymbol{\tau}$.[13,14] That is, fitting the MR model

$$\beta(\tau) = \begin{pmatrix} \mathbf{1}' & \boldsymbol{\tau}' - 0.5 \end{pmatrix} \begin{pmatrix} \beta_m \\ \beta_\tau \end{pmatrix} + \epsilon \tag{8}$$

where β_m is the median effect of the genetic variant, β_τ is the slope coefficient for the change in the median effect with τ, and $\epsilon \in \mathbb{R}^{n \times 1}$ are random errors with the cross-quantile variance-covariance matrix of the estimates under i.i.d errors. This framework provides both location-shift and change in location-shift estimates to further decipher the nature of complex genetic associations.

Simulations: The power to detect unadjusted interactions using CQR and MR was explored using simulations. Equation 1 describes the effects of an interaction between a SNP, G, and a variable, X, on a quantitative trait, Y. Without loss of generality, G was assumed to be biallellic with a MAF, p, under HWE and an additive genetic effect on Y. Moreover, G was encoded such that mean genotype was zero ($-2p$, $1-2p$, or $2-2p$).[11] The total variance of Y was assumed to be 1 and the variance of each component of equation 1 was partitioned accordingly. Specifically, the proportion of the variance (R^2) of Y that was explained by G, X and the interaction between G and X was $R^2_G = 2p(1-p)\beta_2^2$, $R^2_X = \beta_1^2$ and $R^2_{G\times X} = 2p(1-p)\beta_3^2$, respectively. The error term, ϵ, was assumed to have a normal distribution with a mean of 0 and a variance of $1 - R^2_G - R^2_X - R^2_{G\times X}$. Unless otherwise specified, the simulation conditions were MAF $=$ 0.2, N $=$ 10,000, $R^2_G = 0.004$, $R^2_X = 0.25$, and $R^2_{G\times X}$ was varied between 0 and 0.004. When more than one interaction was considered, R^2_X was divided equally between all interaction covariates, while each additional interaction was equal to $R^2_{G\times X}$. All regression models were fitted with Y as the dependent variable and G as the independent variable. CQR models were fitted at every 10$^{\text{th}}$ percentile of the distribution of Y from the 5$^{\text{th}}$ to the 95$^{\text{th}}$ percentiles. A total of 1,000 Markov chain marginal bootstrap (MCMB) replicates were used to compute confidence intervals and the cross-percentile variance-covariance matrix for CQR estimates.[12,15,16] Variability in the CQR estimates of G at these percentiles was modelled using MR, assuming normality, to determine the effects of percentiles on mean CQR estimates. The power to detect interactions at a threshold of p < 0.05 was computed from 1,000 replicates of each simulation condition.

Sample Stratification and Interactions: The analysis of secondary traits (e.g. BMI) collected from case-control studies with disease status (e.g. T2D) as a primary outcome can be prone to artifacts if potential stratification of secondary traits is not addressed.[17] This stems from the fact that secondary traits are often strong risk factors for disease status and can thus be stratified in cases and controls. Since effect alleles of disease-associated SNPs are typically enriched in cases and depleted in controls, the stratification of allele frequencies and secondary traits can correspond. The coinciding stratification of secondary trait distributions and allele frequency distributions may result in spurious associations between these disease-associated SNPs and secondary traits. This phenomenon has also been observed in population-based designs when disease prevalence differs between the sample and general populations.[18] Yaghootkar, et al.,

have recently developed an analytical model relating regression estimate bias to differences between disease prevalence in the sample and general populations.[18] This model described regression estimate bias in the main effects of SNPs as a function of the partitioning of allele frequencies by disease status as well as the partitioning of variance by genotype (i.e. heteroscedasticity). They also extended this description to include regression models fitted with adjustment for disease status and show that the bias persists even after this adjustment.[18] Importantly, when regression models are adjusted for disease status the bias in regression estimates is *not* a function of the partitioning partitioning of variance by genotype.[18] This is critical because it means that while estimates of the main effects of SNPs from CQR models may be affected by sample stratification in the same way as estimates from OLS models, the variation of CQR estimates across the sample distribution is not a function of differences in disease prevalence between sample and general populations. The analytical model presented here is not primarily concerned with main effects of SNPs on continuous outcomes, rather with modelling the *variation* of CQR estimates across the sample outcome distribution.

The effect of sample stratification on the power to detect of unadjusted gene interactions with CQR and MR was assessed in simulations. Consider the disease outcome (Z), the continuous risk factor (Y) and the SNP (G), whose relationship is described using a liability scale disease (probit) model.[18]

$$z_i = \beta_4 g_i + \beta_5 y_i + \varphi_i \tag{9}$$

where the coefficients β_4 and β_5 represent the respective marginal effects of G and Y on Z, φ_i is the random error with $\varphi \sim F_{\varphi}(0, \sigma_{\varphi})$, and y_i is specified in equation 1. Disease status (D) is defined as follows;

$$\alpha = \Phi^{-1}(1 - \pi_0) \tag{10}$$

$$D = \begin{cases} 1 & \text{if } z_i > \alpha \\ 0 & \text{if } z_i \leq \alpha \end{cases} \tag{11}$$

where π_0 is the disease prevalence in the general population. Figure S2A shows a schematic representation of this model. A population of 100,000 individuals was simulated with the following conditions; $\pi_0 = 0.1$ (i.e. population disease prevalence of 10%), MAF $= 0.2$, $R^2_{G[Y]} = 0.004$, $R^2_X = 0.25$, $R^2_{G[Z]} = 0.01$ (equivalent to OR ~ 1.4 for G on D), $R^2_Y = 0.20$ (equivalent to OR ~ 2.5 for Y on D) and $R^2_{G \times X[Y]}$ was varied between 0 and 0.004. A random sample of N $= 10{,}000$ individuals was then drawn from this population with pre-specified proportion of cases (5, 10, 25 and 50%) and then disease adjusted CQR models ($y \sim g + D$) were fitted across the distribution of Y as in simulations above. Variability in the CQR estimates of G at these percentiles was modelled using MR to determine the effects of percentiles on mean CQR estimates. The power to detect interactions at a threshold of p < 0.05 was computed from 1,000 replicates of each simulation condition.

SUPPLEMENTAL REFERENCES

1. Begum, F., Ghosh, D., Tseng, G.C., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Res 40, 3777–3784.

2. Li, A., and Meyre, D. (2013). Challenges in reproducibility of genetic association studies: lessons learned from the obesity field. Int J Obes (Lond) 37, 559–567.

3. Koenker, R. (2005). Quantile Regression (Cambridge University Press).

4. Koenker, R., and Hallock, K. (2001). Quantile regression: An introduction. Journal of Economic Perspectives.

5. Kiserud, T., Piaggio, G., Carroli, G., Widmer, M., Carvalho, J., Neerup Jensen, L., Giordano, D., Cecatti, J.G., Abdel Aleem, H., Talegawkar, S.A., et al. (2017). The World Health Organization Fetal Growth Charts: A Multinational Longitudinal Study of Ultrasound Biometric Measurements and Estimated Fetal Weight. PLoS Med 14, e1002220.

6. Fernández, J.R., Redden, D.T., Pietrobelli, A., and Allison, D.B. (2004). Waist circumference percentiles in nationally representative samples of African-American, European-American, and Mexican-American children and adolescents. J Pediatr 145, 439–444.

7. Wei, Y., Pere, A., Koenker, R., and He, X. (2006). Quantile regression methods for reference growth charts. Statist. Med. 25, 1369–1382.

8. Beyerlein, A. (2014). Quantile Regression–Opportunities and Challenges From a User's Perspective. American Journal of Epidemiology 180, 330–331.

9. Beyerlein, A., Kries, von, R., Ness, A.R., and Ong, K.K. (2011). Genetic markers of obesity risk: stronger associations with body composition in overweight compared to normal-weight children. PLoS ONE 6, e19057.

10. Mitchell, J.A., Hakonarson, H., Rebbeck, T.R., and Grant, S.F.A. (2013). Obesity-susceptibility loci and the tails of the pediatric BMI distribution. Obesity 21, 1256–1260.

11. Paré, G., Cook, N.R., Ridker, P.M., and Chasman, D.I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. PLoS Genet 6, e1000981.

12. Koenker, R., and Bassett, G., Jr (1982). Robust tests for heteroscedasticity based on regression quantiles. Econometrica 50, 43–61.

13. Thompson, S.G., and Higgins, J.P.T. (2002). How should meta-regression analyses be undertaken and interpreted? Statist. Med. 21, 1559–1573.

14. Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R. (2009). Meta-Regression, in Introduction to Meta-Analysis (Chichester, UK: John Wiley & Sons,

Ltd).

15. He, X., and Hu, F. (2002). Markov chain marginal bootstrap. Journal of the American Statistical Association 97, 783–795.

16. Kocherginsky, M., He, X., and Mu, Y. (2005). Practical Confidence Intervals for Regression Quantiles. Journal of Computational and Graphical Statistics 14, 41–55.

17. Monsees, G.M., Tamimi, R.M., and Kraft, P. (2009). Genome-wide association scans for secondary traits using case-control samples. Genet. Epidemiol. 33, 717–728.

18. Yaghootkar, H., Bancks, M.P., Jones, S.E., McDaid, A., Beaumont, R., Donnelly, L., Wood, A.R., Campbell, A., Tyrrell, J., Hocking, L.J., et al. (2017). Quantifying the extent to which index event biases influence large genetic association studies. Hum Mol Genet 26, 1018–1030.

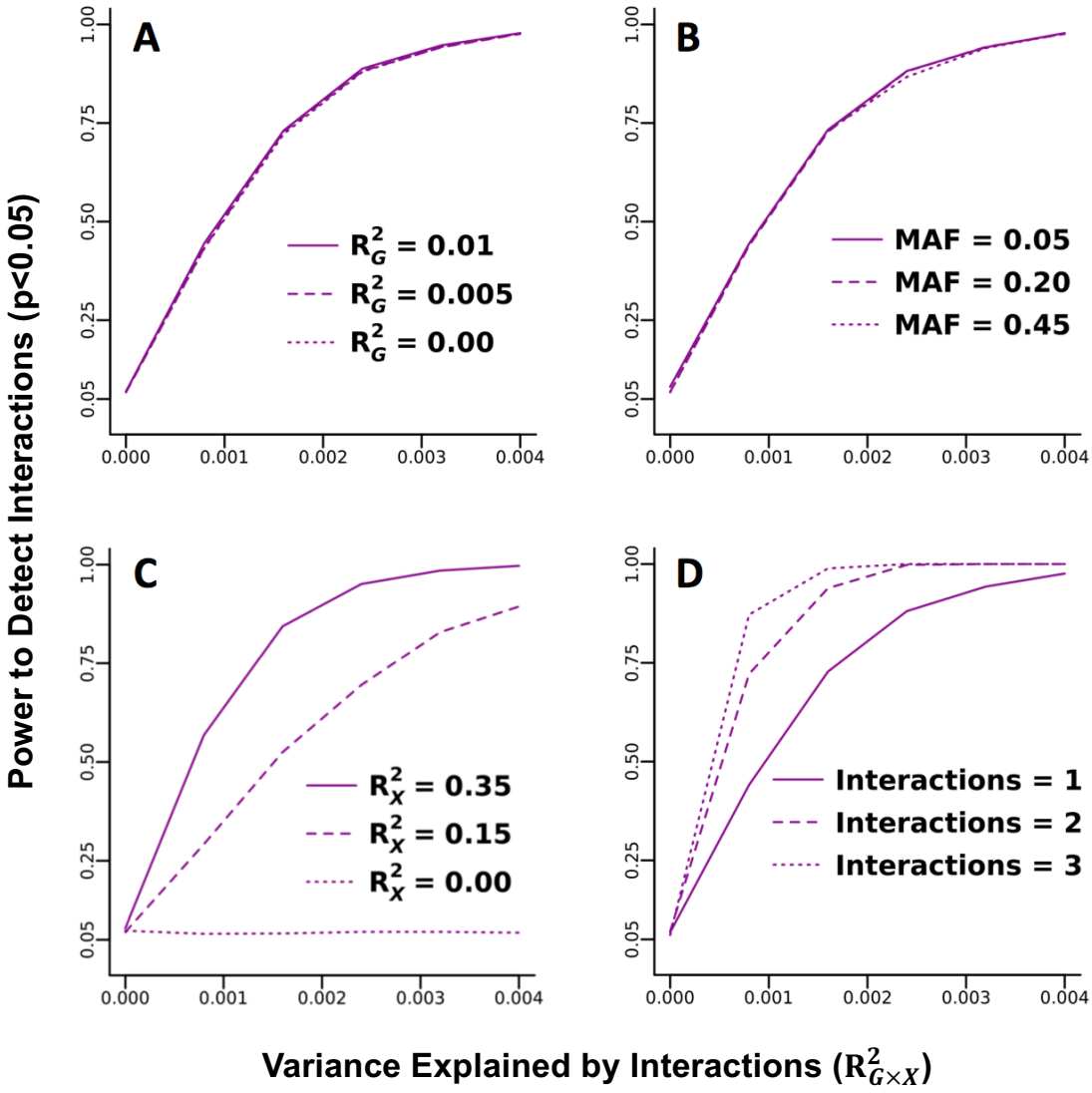


Figure S1: Simulation study of the power to detect unadjusted interactions using conditional quantile regression (CQR) and meta-regression (MR). The power to detect unadjusted interactions between a SNP (*G*) and a continuous variable (*X*) was simulated in a sample of 10,000 individuals. Unless otherwise indicated, the simulation conditions were minor allele frequency (MAF) = 0.2, variance explained by *G* (R^2_G) = 0.004, variance explained by *X* (R^2_X) = 0.25, and the variance explained by the interaction between *G* and *X* ($R^2_{G\times X}$) was varied between 0 and 0.004. CQR models were fitted at every 10th percentile of the distribution of *Y* from the 5th to the 95th percentiles and MR was used to model the relationship between variation in CQR estimates and the *Y* percentiles. The power to detect unadjusted interactions at a threshold of $p < 0.05$ was computed from 1,000 replicates of each simulation condition and plotted against the value of $R^2_{G\times X}$. The power to detect interactions at different values of R^2_G MAF, R^2_X and the number of interactions was investigated (A, B, C and D, respectively). When more than one interaction was considered, R^2_X was divided equally between all interaction covariates, while each additional interaction was equal to $R^2_{G\times X}$. Overall the power to detect unadjusted interactions was not affected by the main effects of *G* or the MAF, but was enhanced by the main effects of *X* and the number of interactions.

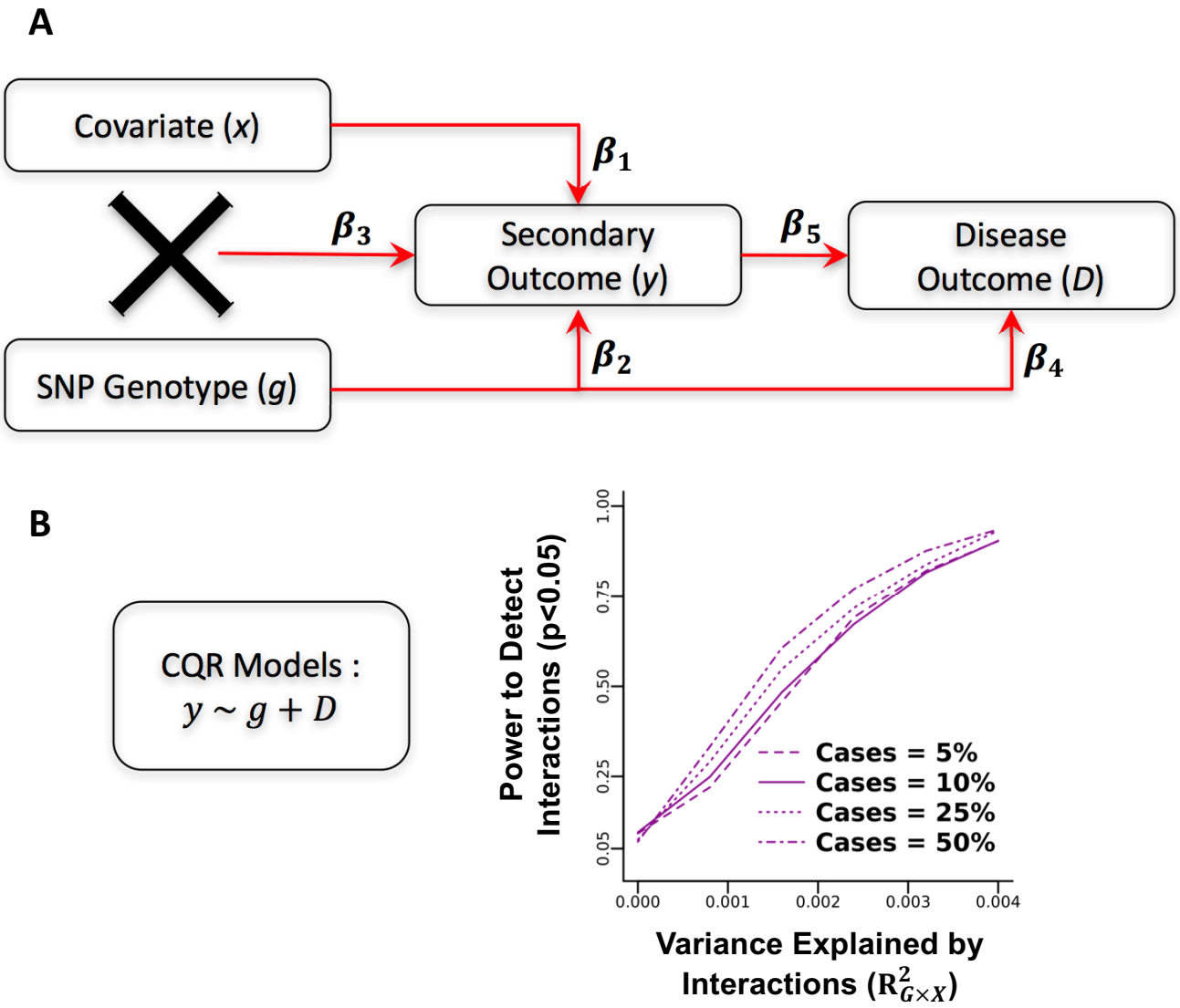
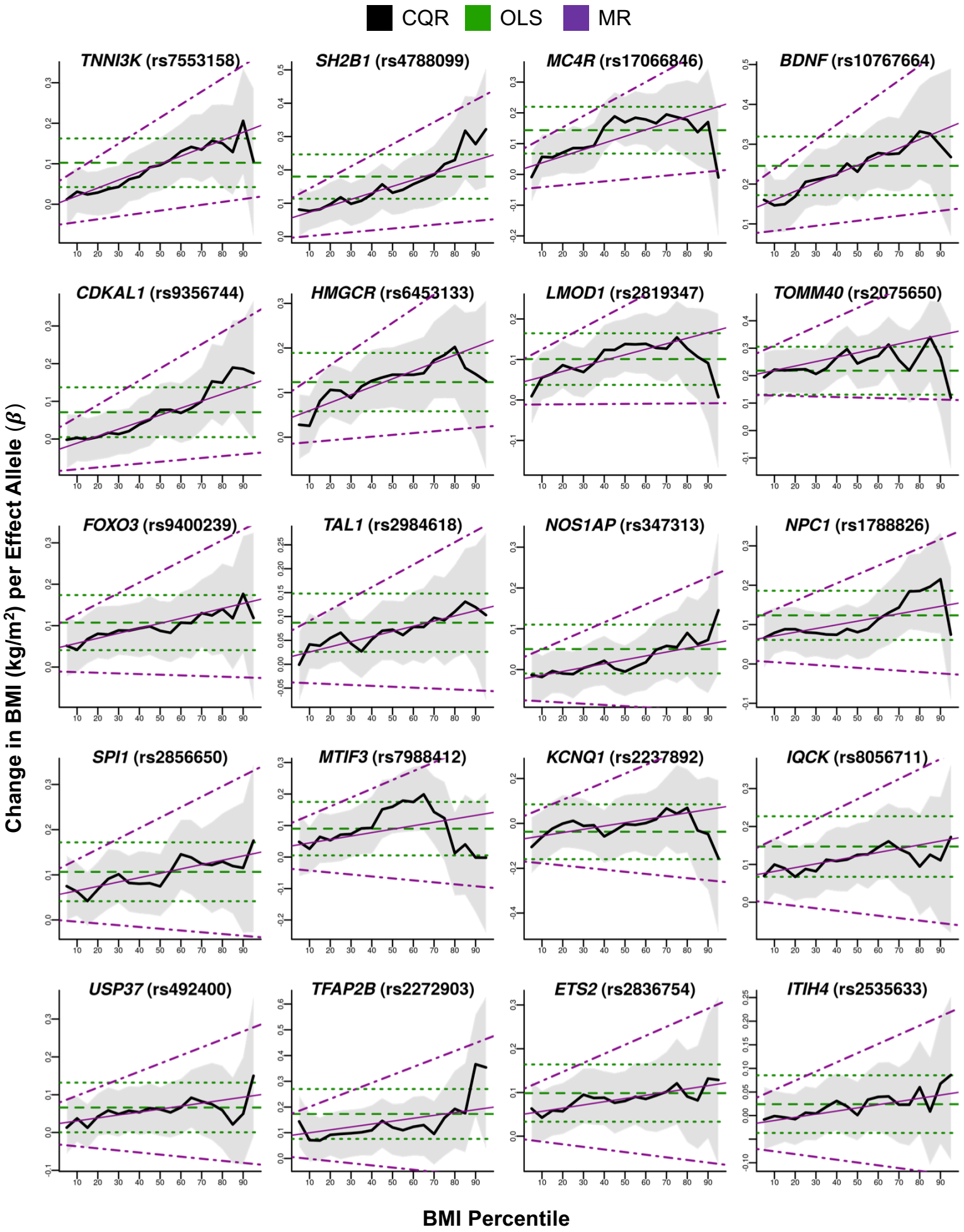


Figure S2: Sample stratification and the detection of unadjusted interactions in simulations. (A) A schematic representation of the model described by equations 1, 9, 10 and 11 in Appendix A. (B) Investigating the effects of sample stratification on the power to detect unadjusted interactions using conditional quantile regression (CQR) and meta-regression (MR) in a simulation study. The simulation conditions were minor allele frequency (MAF) = 0.2, variance of *Z* explained by *G* ($R^2_{G[Z]}$) = 0.01 (equivalent to OR ~ 1.4 of *G* on *D*), variance *Z* explained by *Y* (R^2_Z) = 0.2 (equivalent to OR ~ 2.5 of *Y* on *D*), variance of *Y* explained by *G* ($R^2_{G[Y]}$) = 0.004, variance of *Y* explained by *X* (R^2_X) = 0.25 and the variance of *Y* explained by the interaction between *G* and *X* ($R^2_{G\times X}$) was varied between 0 and 0.004. A population of 100,000 individuals was generated with disease prevalence (π_0) = 10%. A sample population of 10,000 individuals with pre-specified proportion of cases was randomly selected from this population. The power to detect unadjusted interactions between the SNP (*G*) and the continuous variable (*X*) in this sample was computed and plotted as in Figure S1, except that CQR models were adjusted for disease status (*D*). Overall the power to detect unadjusted interactions was not affected by sample stratification when CQR models were adjusted for disease status.


CQR
OLS
MR
TNNI3K (rs7553158)
SH2B1 (rs4788099)
MC4R (rs17066846)
BDNF (rs10767664)
CDKAL1 (rs9356744)
HMGCR (rs6453133)
LMOD1 (rs2819347)
TOMM40 (rs2075650)
FOXO3 (rs9400239)
TAL1 (rs2984618)
NOS1AP (rs347313)
NPC1 (rs1788826)
SPI1 (rs2856650)
MTIF3 (rs7988412)
KCNQ1 (rs2237892)
IQCK (rs8056711)
USP37 (rs492400)
TFAP2B (rs2272903)
ETS2 (rs2836754)
ITIH4 (rs2535633)
Change in BMI (kg/m2) per Effect Allele (β)
BMI Percentile

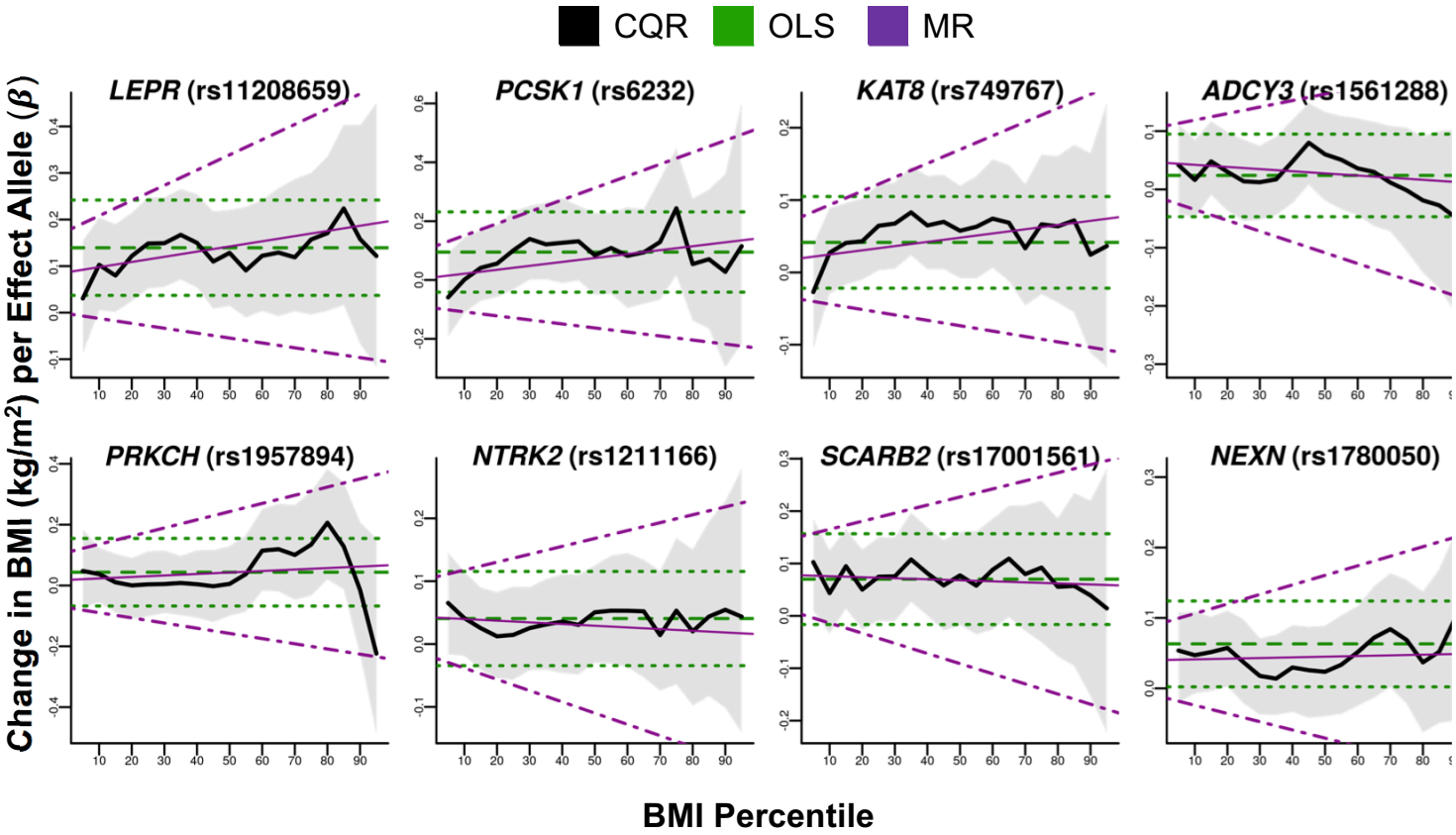
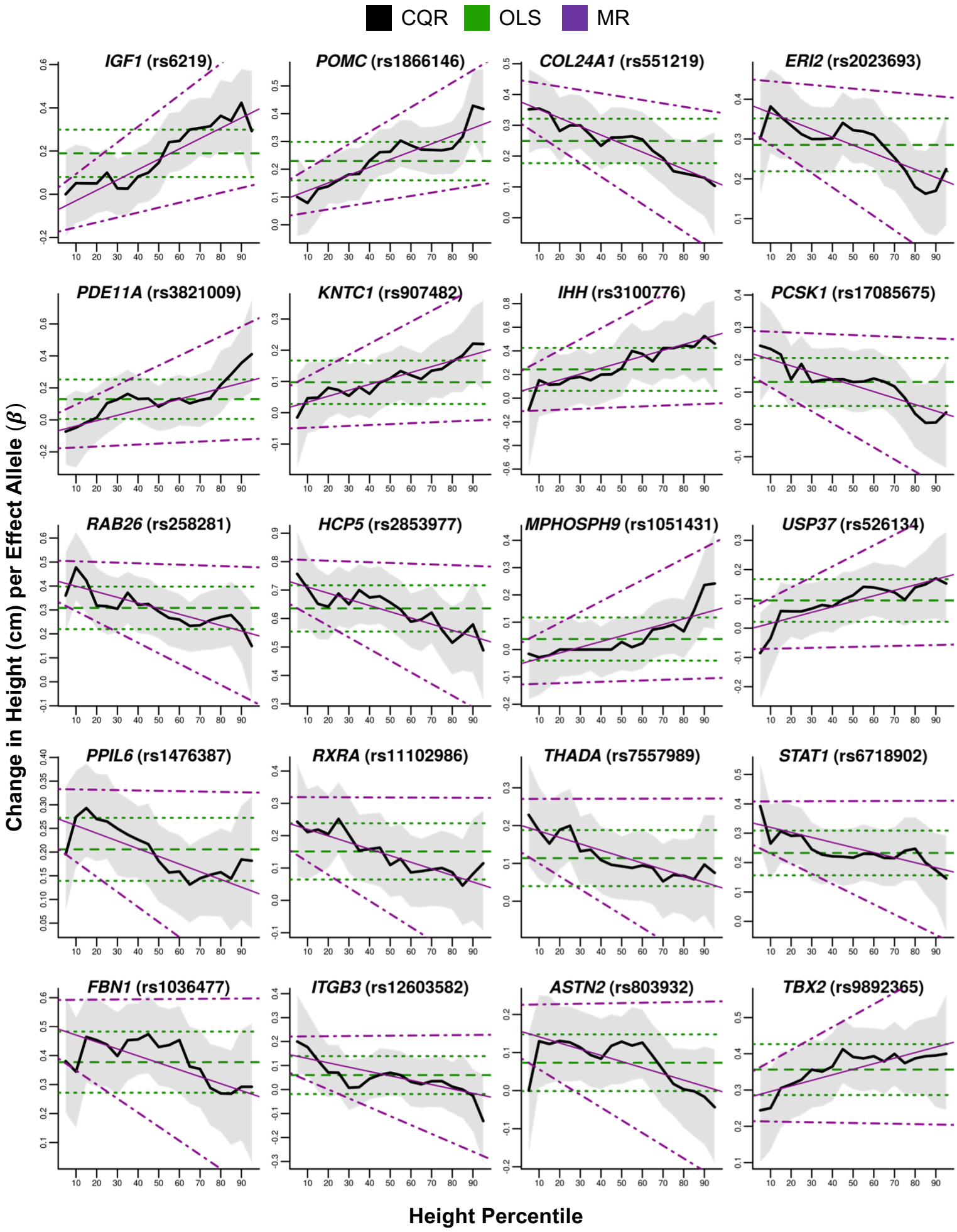
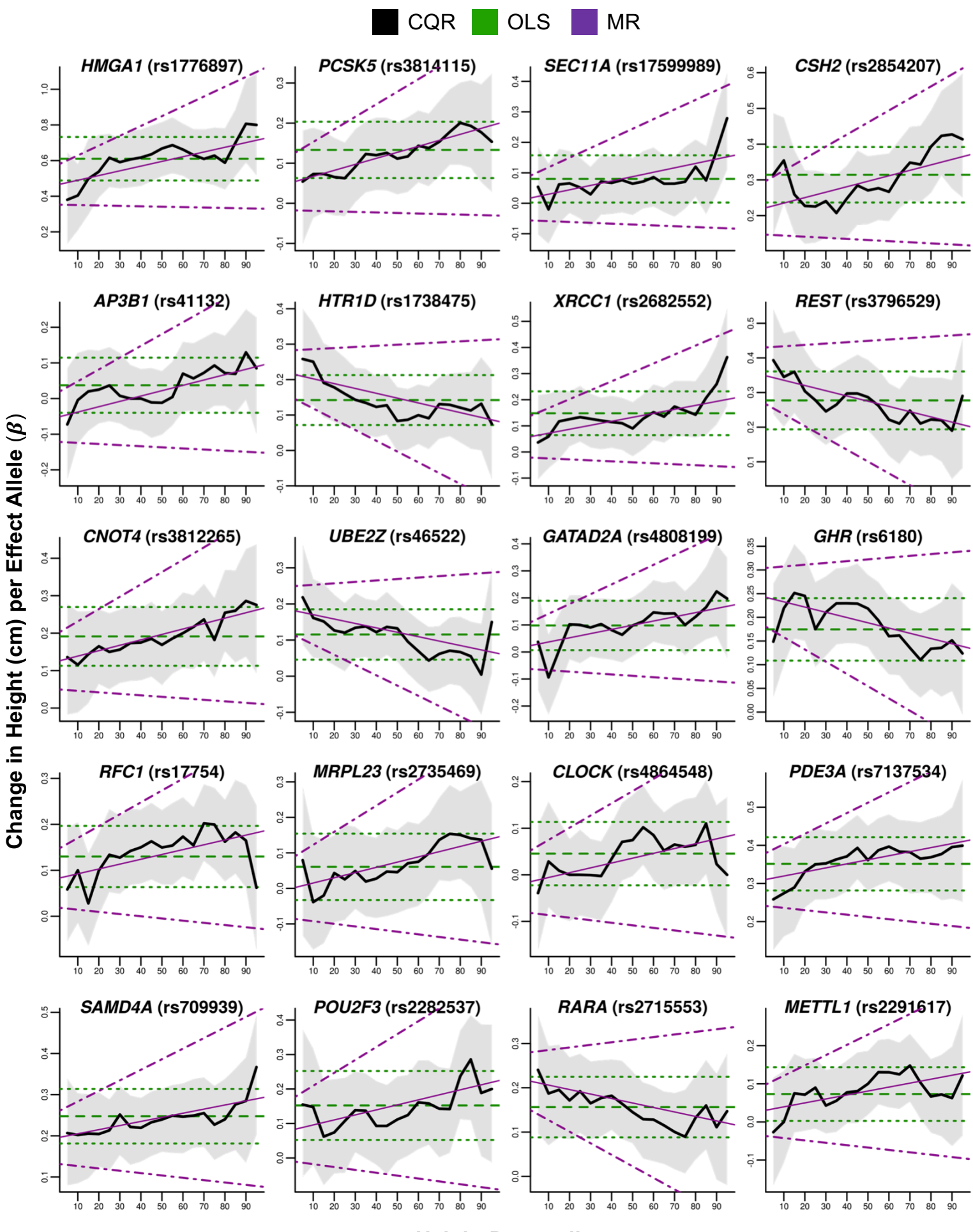
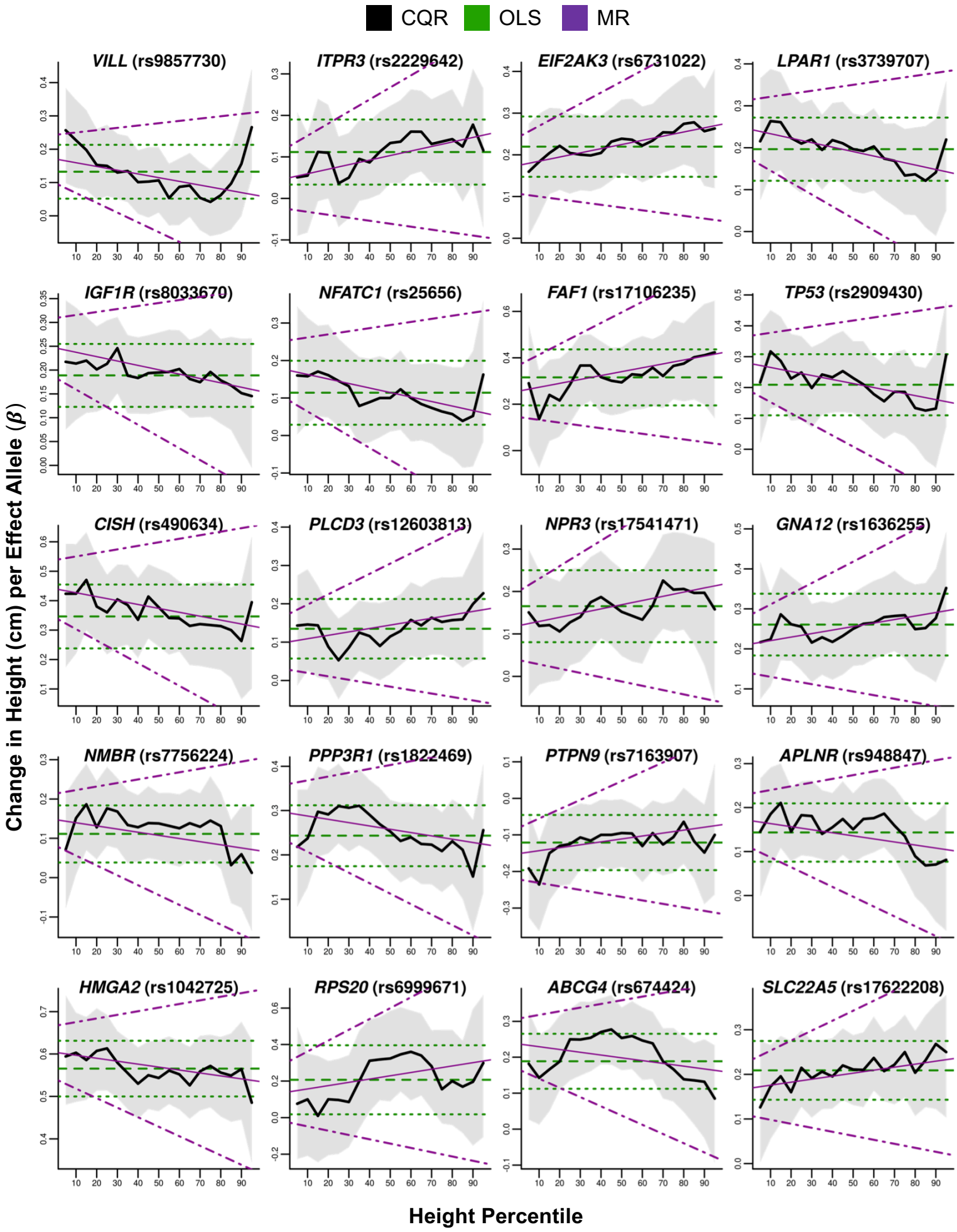
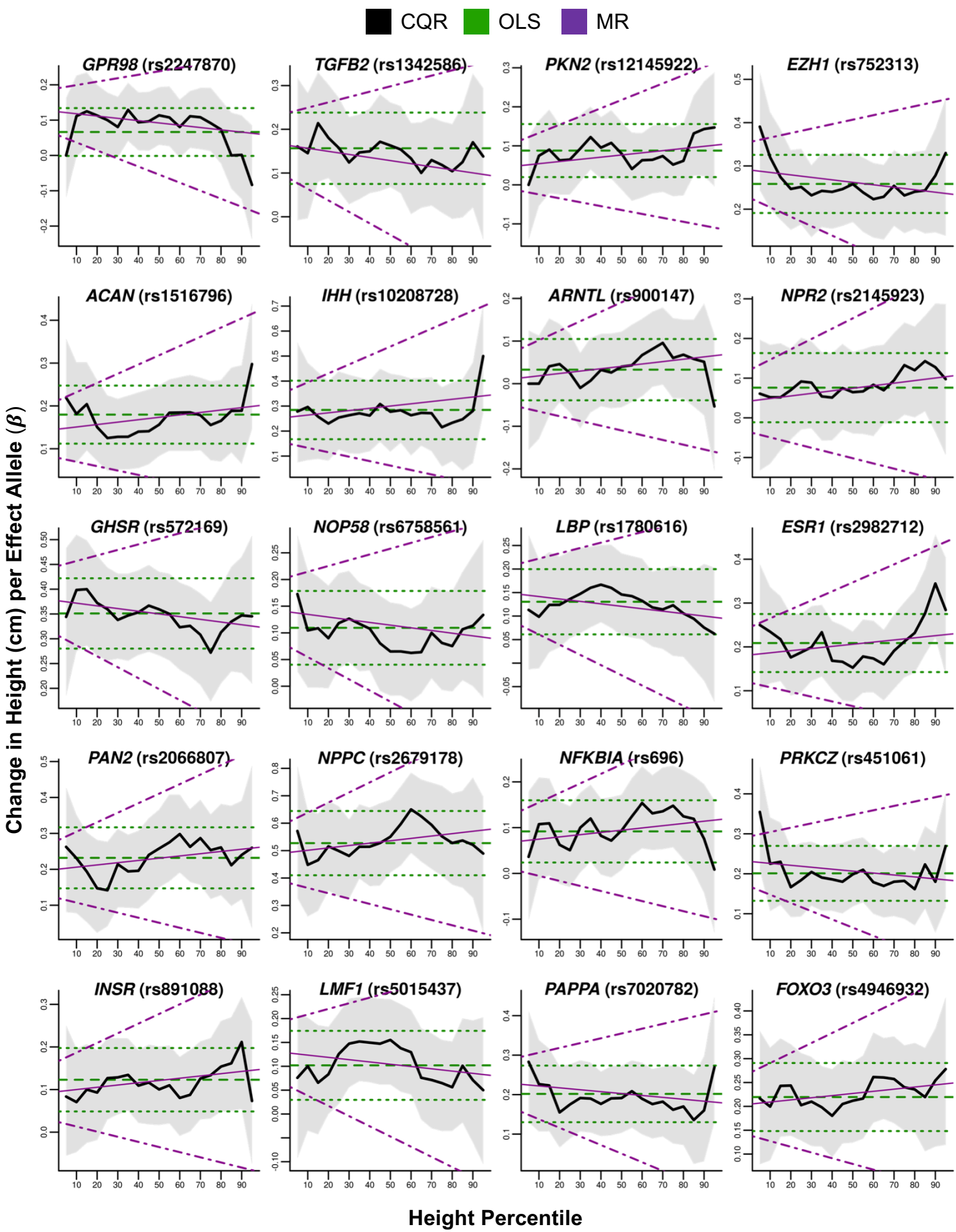


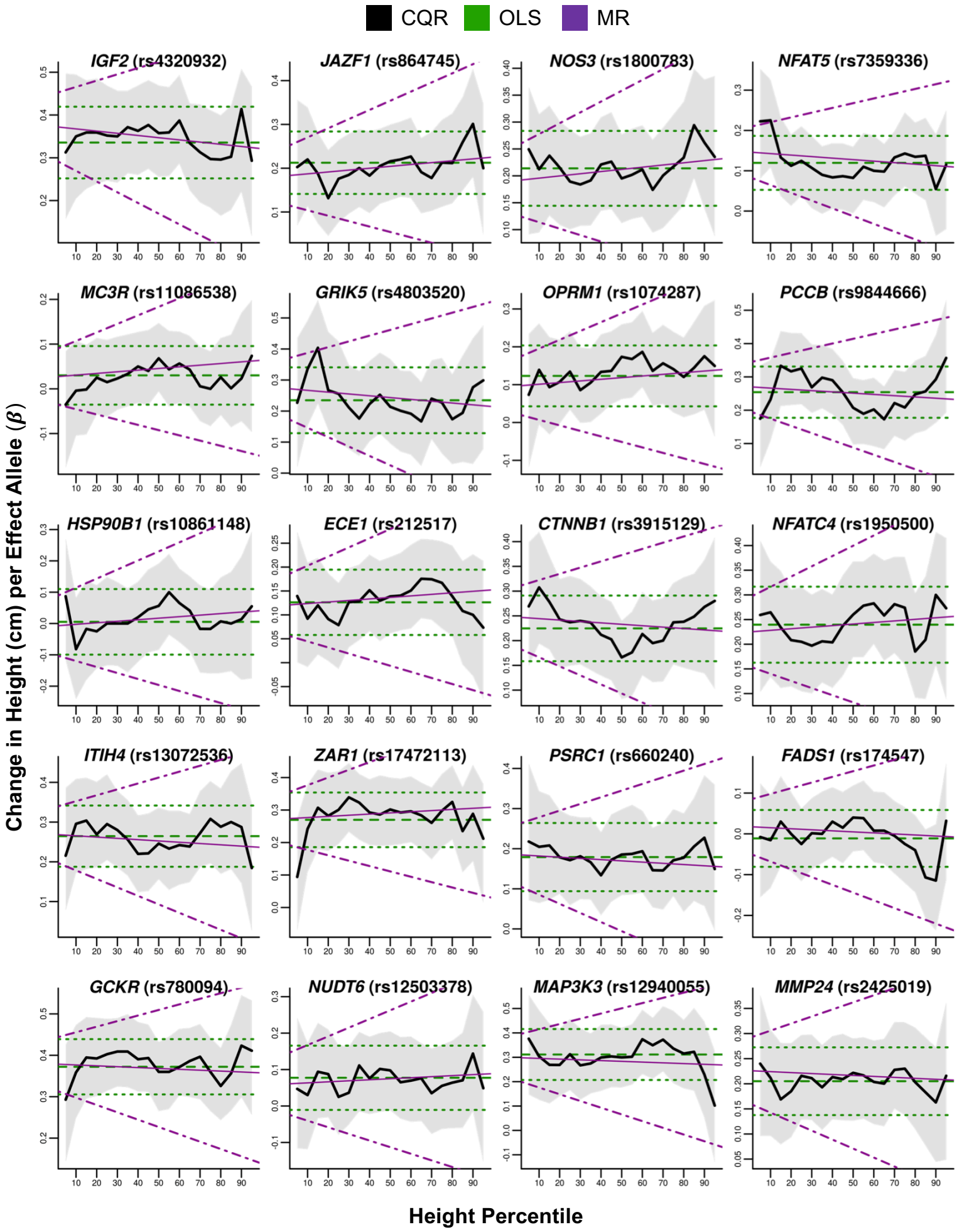
Figure S3: The effects of BMI/obesity-associated SNPs across the sample BMI distribution (continued). As in Figure 2, estimates of the change in BMI per effect allele (β_{CQR} , kg/m^2 per Effect Allele) from conditional quantile regression (CQR) models of BMI/obesity-associated SNPs was plotted against the BMI percentile (thick-black line) along with the 95% confidence intervals (shaded-grey region). The results from ordinary least square (OLS) (β_{OLS} , kg/m^2 per Effect Allele, horizontal-dashed-green line) and the 95% confidence intervals (horizontal-dashed-green lines) were also plotted for comparison. The change in CQR estimates across BMI percentiles was modelled using meta-regression (MR) and estimates from MR (β_{MR} , kg/m^2 per Effect Allele per BMI Percentile, thin-magenta line) and the 95\% confidence intervals (dotdashed-magenta lines) were plotted. MR analysis did not detect significant ($p < 1.32x10^{-03}$) increases in the effects of these SNPs across the sample BMI distribution.

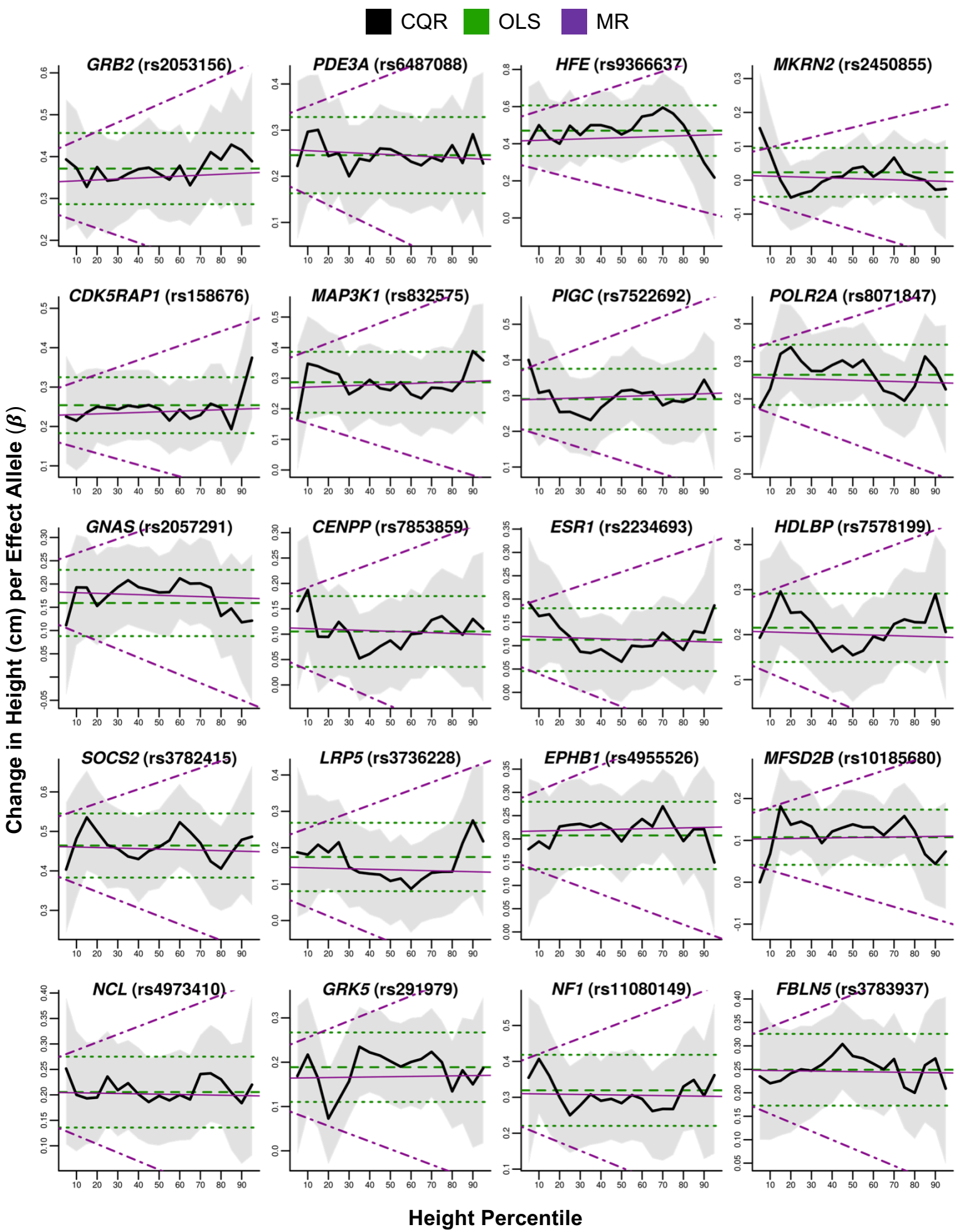

CQR
OLS
MR
IGF1 (rs6219)
POMC (rs1866146)
COL24A1 (rs551219)
ERI2 (rs2023693)
PDE11A (rs3821009)
KNTC1 (rs907482)
IHH (rs3100776)
PCSK1 (rs17085675)
RAB26 (rs258281)
HCP5 (rs2853977)
MPHOSPH9 (rs1051431)
USP37 (rs526134)
PPIL6 (rs1476387)
RXRA (rs11102986)
THADA (rs7557989)
STAT1 (rs6718902)
FBN1 (rs1036477)
ITGB3 (rs12603582)
ASTN2 (rs803932)
TBX2 (rs9892365)
Change in Height (cm) per Effect Allele (β)
Height Percentile


CQR
OLS
MR
HMGA1 (rs1776897)
PCSK5 (rs3814115)
SEC11A (rs17599989)
CSH2 (rs2854207)
AP3B1 (rs41132)
HTR1D (rs1738475)
XRCC1 (rs2682552)
REST (rs3796529)
CNOT4 (rs3812265)
UBE2Z (rs46522)
GATAD2A (rs4808199)
GHR (rs6180)
RFC1 (rs17754)
MRPL23 (rs2735469)
CLOCK (rs4864548)
PDE3A (rs7137534)
SAMD4A (rs709939)
POU2F3 (rs2282537)
RARA (rs2715553)
METTL1 (rs2291617)
Change in Height (cm) per Effect Allele (β)
Height Percentile


CQR
OLS
MR
VILL (rs9857730)
ITPR3 (rs2229642)
EIF2AK3 (rs6731022)
LPAR1 (rs3739707)
IGF1R (rs8033670)
NFATC1 (rs25656)
FAF1 (rs17106235)
TP53 (rs2909430)
CISH (rs490634)
PLCD3 (rs12603813)
NPR3 (rs17541471)
GNA12 (rs1636255)
NMBR (rs7756224)
PPP3R1 (rs1822469)
PTPN9 (rs7163907)
APLNR (rs948847)
HMGA2 (rs1042725)
RPS20 (rs6999671)
ABCG4 (rs674424)
SLC22A5 (rs17622208)
Change in Height (cm) per Effect Allele (β)
Height Percentile


CQR
OLS
MR
GPR98 (rs2247870)
TGFB2 (rs1342586)
PKN2 (rs12145922)
EZH1 (rs752313)
ACAN (rs1516796)
IHH (rs10208728)
ARNTL (rs900147)
NPR2 (rs2145923)
GHSR (rs572169)
NOP58 (rs6758561)
LBP (rs1780616)
ESR1 (rs2982712)
PAN2 (rs2066807)
NPPC (rs2679178)
NFKBIA (rs696)
PRKCZ (rs451061)
INSR (rs891088)
LMF1 (rs5015437)
PAPPA (rs7020782)
FOXO3 (rs4946932)
Change in Height (cm) per Effect Allele (β)
Height Percentile


CQR
OLS
MR
IGF2 (rs4320932)
JAZF1 (rs864745)
NOS3 (rs1800783)
NFAT5 (rs7359336)
MC3R (rs11086538)
GRIK5 (rs4803520)
OPRM1 (rs1074287)
PCCB (rs9844666)
HSP90B1 (rs10861148)
ECE1 (rs212517)
CTNNB1 (rs3915129)
NFATC4 (rs1950500)
ITIH4 (rs13072536)
ZAR1 (rs17472113)
PSRC1 (rs660240)
FADS1 (rs174547)
GCKR (rs780094)
NUDT6 (rs12503378)
MAP3K3 (rs12940055)
MMP24 (rs2425019)
Change in Height (cm) per Effect Allele (β)
Height Percentile


CQR
OLS
MR
GRB2 (rs2053156)
PDE3A (rs6487088)
HFE (rs9366637)
MKRN2 (rs2450855)
CDK5RAP1 (rs158676)
MAP3K1 (rs832575)
PIGC (rs7522692)
POLR2A (rs8071847)
GNAS (rs2057291)
CENPP (rs7853859)
ESR1 (rs2234693)
HDLBP (rs7578199)
SOCS2 (rs3782415)
LRP5 (rs3736228)
EPHB1 (rs4955526)
MFSD2B (rs10185680)
NCL (rs4973410)
GRK5 (rs291979)
NF1 (rs11080149)
FBLN5 (rs3783937)
Change in Height (cm) per Effect Allele (β)
Height Percentile

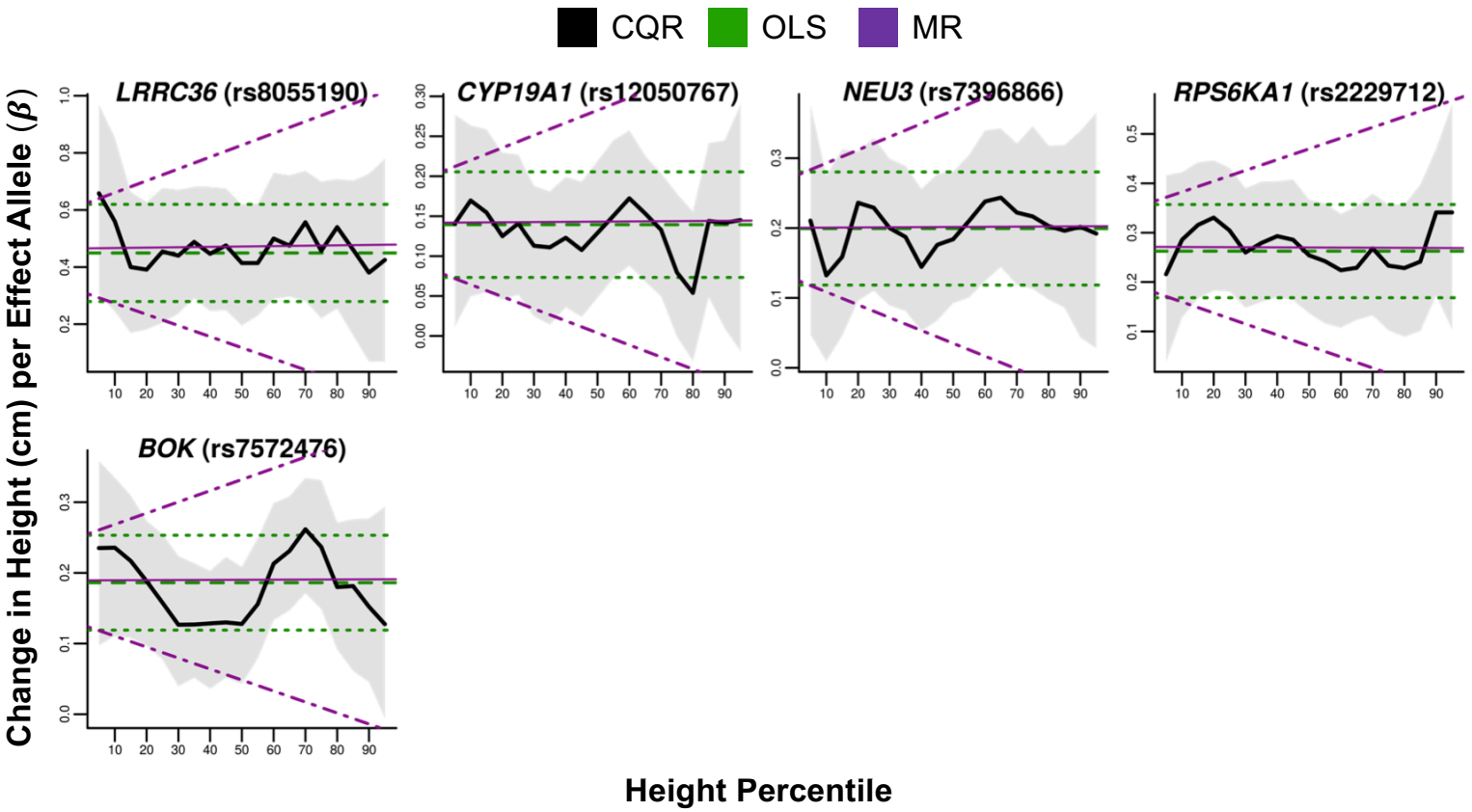
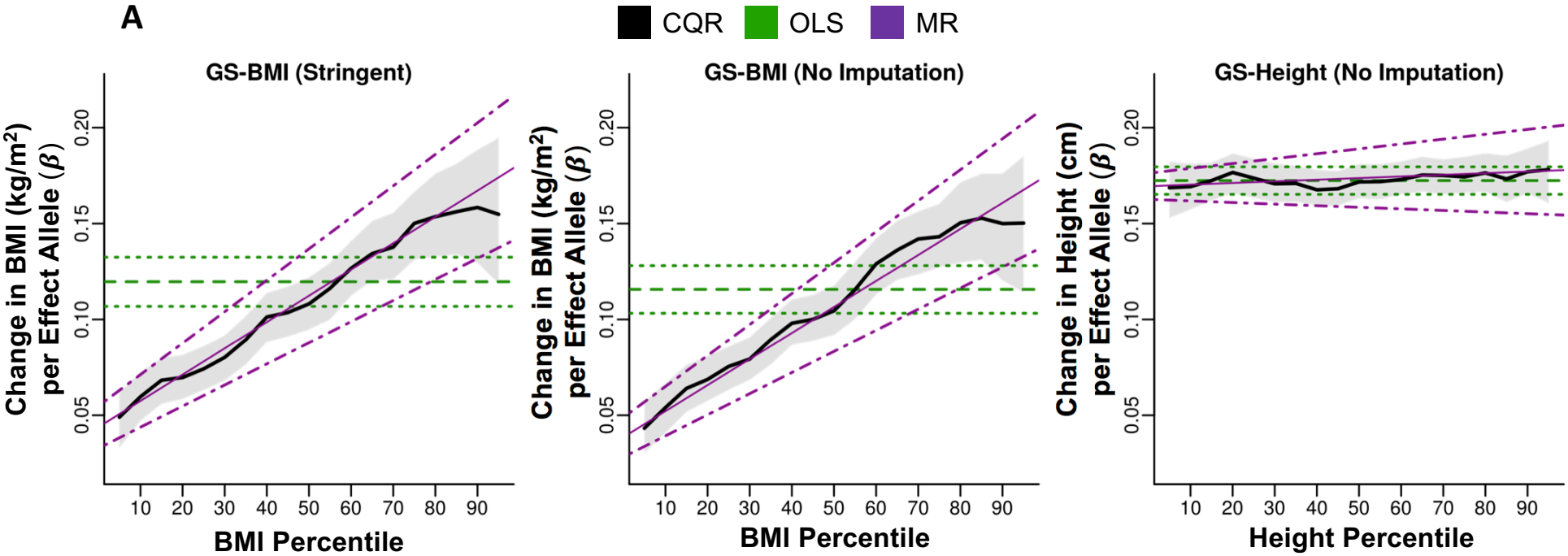


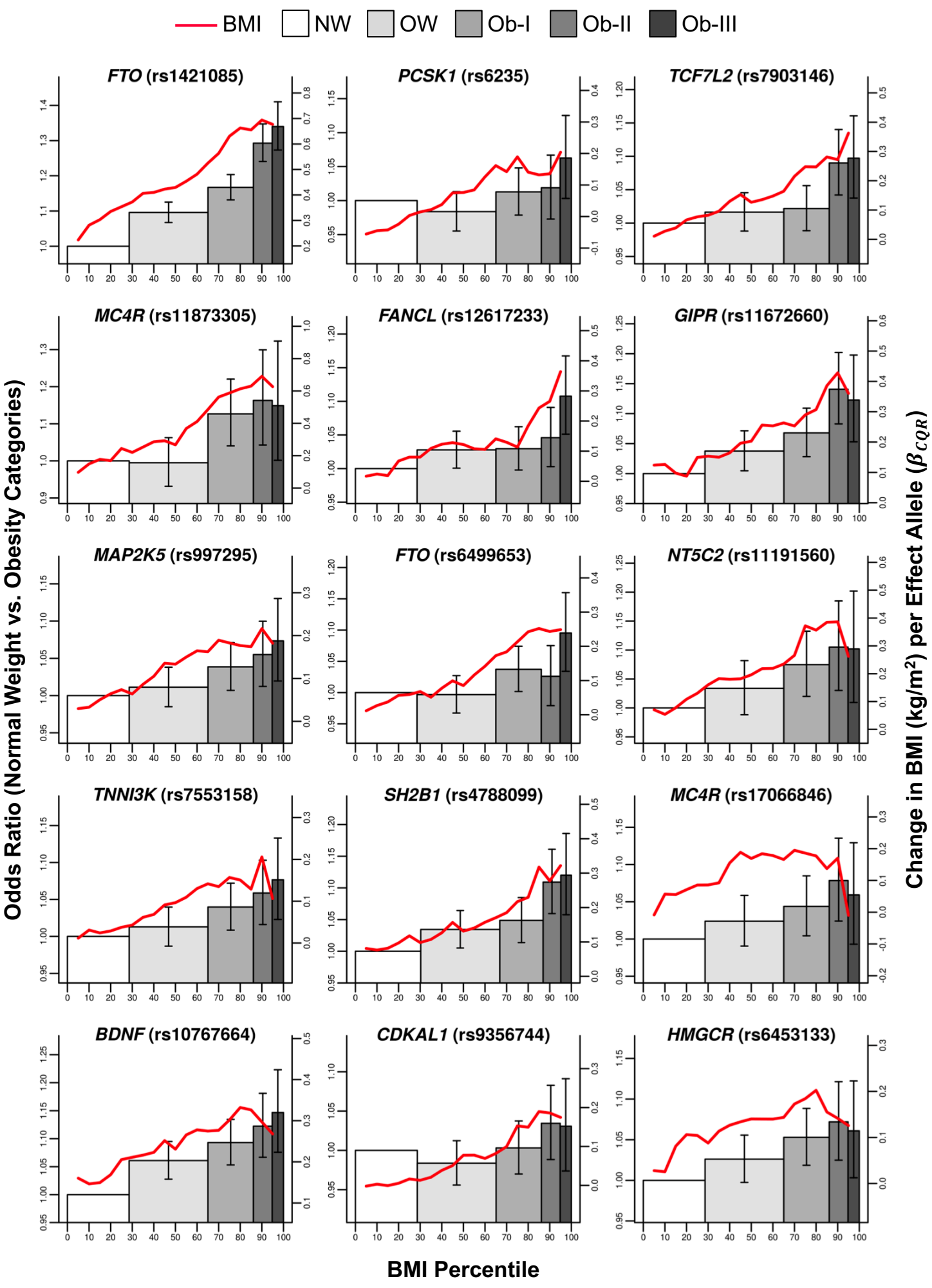
Figure S4: The effects of height-associated SNPs across the distribution of height. Conditional quantile regression (CQR) models of height-associated SNPs were fitted every 5th percentile of height and adjusted for age, sex and study. Estimates of the change in height per effect allele (β_{CQR} , cm per Effect Allele) from these models was plotted against the height percentile (thick-black line) along with the 95% confidence intervals (shaded-grey region). The results from ordinary least square (OLS) models (β_{OLS} , cm per Effect Allele, horizontal-dashed-green line) and the 95% confidence intervals (horizontal-dotted-green lines) were also plotted for comparison. The change in CQR estimates across height percentiles was modelled using meta-regression (MR) and estimates from MR (β_{MR} , cm per Effect Allele per Height Percentile, thin-magenta line) and the 95% confidence intervals (dotdashed-magenta lines) were plotted.

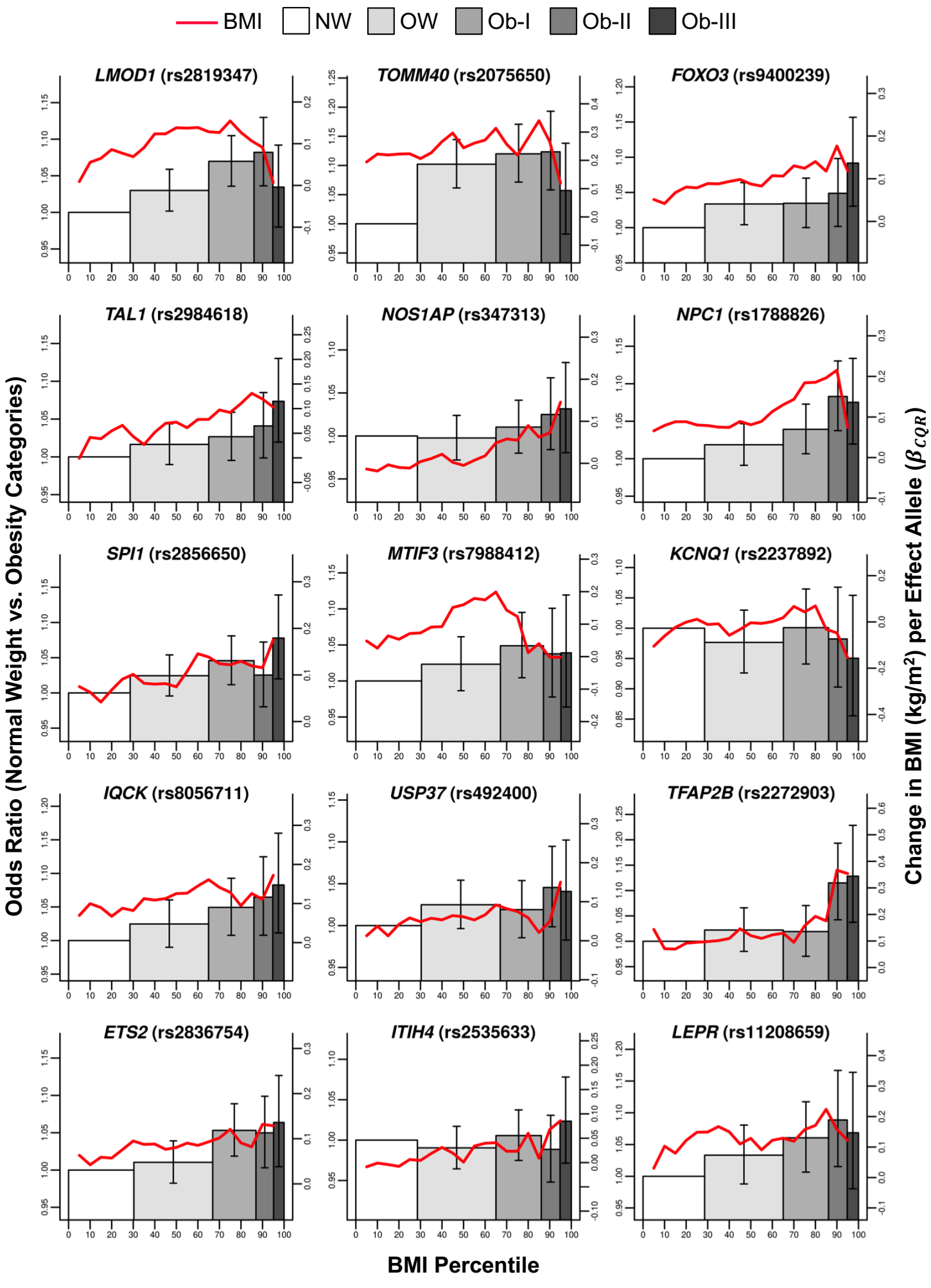


B

SNP	OLS Models		Meta-Regression of CQR Models		
	β_{OLS} [95%CI]	p-value	RI_{50}	β_{MR} [95%CI]	p-value
GS-BMI (Stringent)	0.120 [0.107, 0.132]	2.02E-74	0.112	0.137 [0.110, 0.164]	2.18E-23 *
GS-BMI (No Imputation)	0.116 [0.103, 0.128]	5.80E-74	0.106	0.136 [0.110, 0.161]	3.93E-25 *
GS-Height (No Imputation)	0.172 [0.165, 0.180]	<2.2E-308	0.174	0.009 [-0.008, 0.025]	0.323

Figure S5: Sensitivity analysis of GS Results. (A) CQR models of GS-BMI (Stringent), GS-BMI (No Imputation) and GS-Height (No Imputation) fitted as in Figure 2 and plotted against respective outcome percentiles. The thick-black line is the estimated change in each trait per effect allele (BMI, β_{CQR}, kg/m^2 per Effect Allele; Height, β_{CQR}, cm per Effect Allele) and shaded-grey region represents the 95% confidence intervals. Also plotted are the OLS regression estimates (BMI, β_{OLS} in kg/m^2 per Effect Allele; Height, β_{OLS}, cm per Effect Allele, horizontal-dashed-green line) and 95% confidence intervals (horizontal-dotted-green lines). The change in CQR estimates across outcome percentiles was modeled using meta-regression (MR). Estimates from MR (BMI, β_{MR}, kg/m^2 per Effect Allele per BMI Percentile; Height, β_{MR}, cm per Effect Allele per Height Percentile; thin-magenta line) and the 95% confidence intervals (dotdashed-magenta lines) were also plotted. (B) The results from OLS and MR modelling of GS-BMI (Stringent), GS-BMI (No Imputation) and GS-Height (No Imputation). (*) denotes statistical significance, RI_{50} is the re-centered intercept of the MR models and 95%CI are the 95% confidence intervals.


BMI
NW
OW
Ob-I
Ob-II
Ob-III
FTO (rs1421085)
PCSK1 (rs6235)
TCF7L2 (rs7903146)
MC4R (rs11873305)
FANCL (rs12617233)
GIPR (rs11672660)
MAP2K5 (rs997295)
FTO (rs6499653)
NT5C2 (rs11191560)
TNNI3K (rs7553158)
SH2B1 (rs4788099)
MC4R (rs17066846)
BDNF (rs10767664)
CDKAL1 (rs9356744)
HMGCR (rs6453133)
Odds Ratio (Normal Weight vs. Obesity Categories)
Change in BMI (kg/m²) per Effect Allele (β_{CQR})
BMI Percentile


BMI
NW
OW
Ob-I
Ob-II
Ob-III
LMOD1 (rs2819347)
TOMM40 (rs2075650)
FOXO3 (rs9400239)
TAL1 (rs2984618)
NOS1AP (rs347313)
NPC1 (rs1788826)
SPI1 (rs2856650)
MTIF3 (rs7988412)
KCNQ1 (rs2237892)
IQCK (rs8056711)
USP37 (rs492400)
TFAP2B (rs2272903)
ETS2 (rs2836754)
ITIH4 (rs2535633)
LEPR (rs11208659)
Odds Ratio (Normal Weight vs. Obesity Categories)
Change in BMI (kg/m²) per Effect Allele (β_{CQR})
BMI Percentile

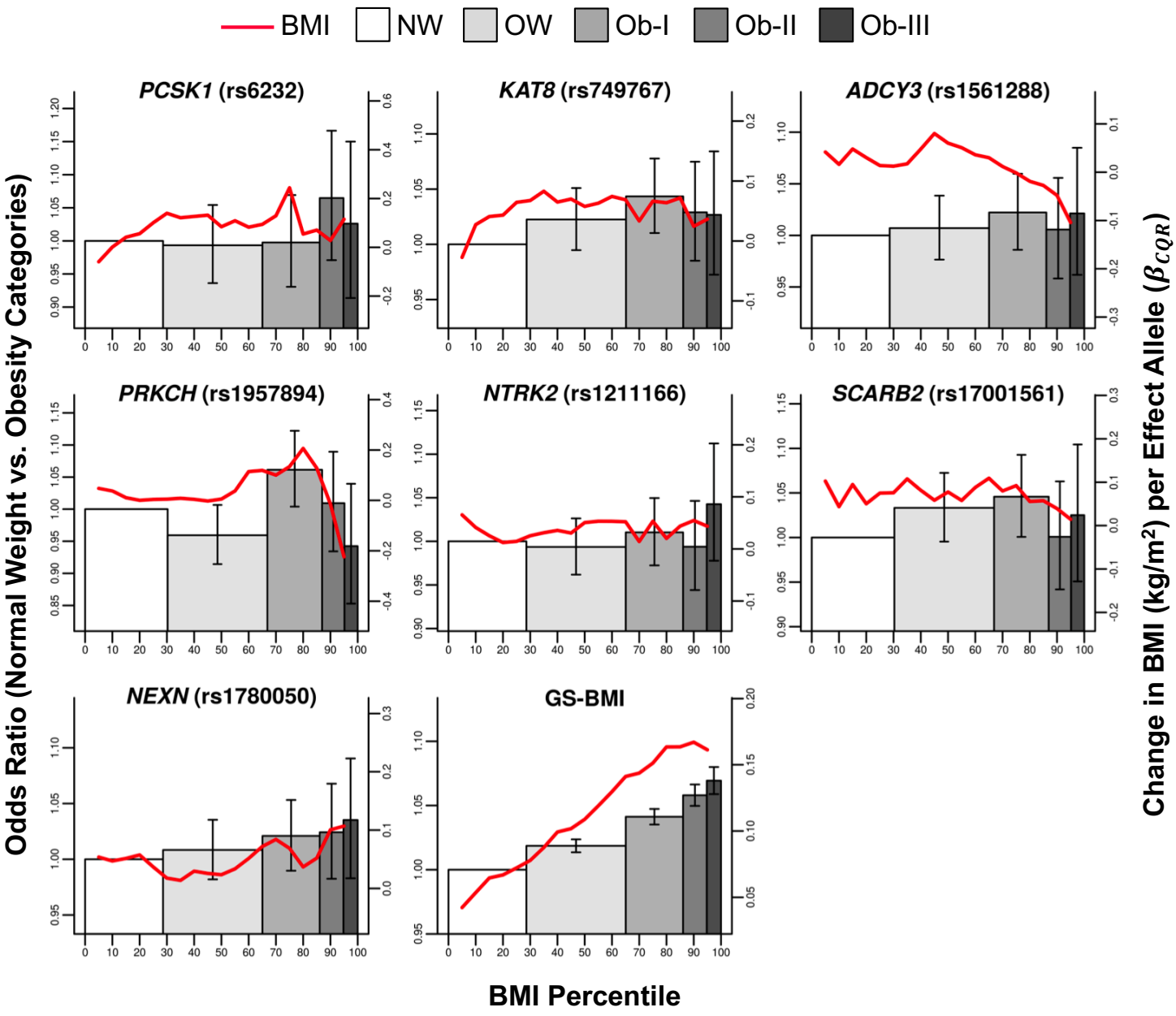


Figure S6: Comparing patterns from subgroup analysis and conditional quantile regression (CQR). BMI was divided into BMI categories, and the effects of each SNP on the risk of overweight (OW), obesity class I (Ob-I), class II (Ob-II) and class III (Ob-III) relative to normal weight (NW) were tested using logistic regression. Models were adjusted for age, age-squared, sex and study. Bar plots of the odds ratio (OR, left axis) for these categories were plotted and bar widths were defined by the percentile cut-offs of each category. Error bars correspond to the 95% confidence intervals. These bar plots were then overlaid with the results from similarly adjusted CQR models (thick-red line, right axis). The patterns from subgroup analysis correspond closely to those from CQR.

Table S1: Subject characteristics. Subject characteristics of the studies included in the analysis of BMI and height; sample size (N), height (mean ± sd), BMI (mean ± sd), age (mean ± sd), the proportion of women, the proportion with diabetes, and the proportions of BMI categories including normal weight (NW), overweight (OW), and obesity classes I (Ob-I), II (Ob-II) and III (Ob-III); within each study and overall are presented.

Study	N	Age (yrs)	Women	Height (cm)	BMI (kg/m^2)	Diabetes	BMI Categories (NW/OW/Ob-I/Ob-II/Ob-III)
ARIC CARe	10094	57.02 ± 6.25	53.40%	168.24 ± 9.47	27.50 ± 5.00	20.50%	33.6/40.8/17.9/5.4/2.3%
CARDIA CARe	1513	40.06 ± 4.20	53.20%	171.45 ± 9.20	26.98 ± 5.83	7.90%	43.1/33.8/13.7/5.9/3.4%
CHS CARe	4194	75.71 ± 5.73	56.40%	164.82 ± 9.50	26.44 ± 4.51	25.60%	40.2/41.6/14.1/3.1/1.1%
Framingham CARe	2222	48.00 ± 13.06	66.00%	166.70 ± 9.43	26.54 ± 5.11	14.90%	43.6/36.4/13.5/4.5/2.0%
MESA CARe	2435	65.47 ± 10.18	52.30%	168.18 ± 9.65	27.87 ± 5.08	14.20%	30.6/40.9/19.5/6.7/2.3%
COPDGene	3793	60.24 ± 8.82	50.20%	169.73 ± 9.51	29.12 ± 5.98	12.30%	26.1/36.5/22.3/9.3/5.8%
eMERGE II	24584	62.12 ± 15.14	54.90%	169.50 ± 10.06	29.51 ± 6.82	17.00%	26.2/34.8/21.4/10.1/7.5%
EpiDREAM	9148	55.10 ± 10.79	60.70%	166.93 ± 9.44	30.67 ± 6.15	14.70%	16.1/35.7/28.0/12.4/7.8%
MOPMAP (WHI)	2197	66.23 ± 6.87	100%	162.19 ± 5.81	28.93 ± 5.76	5.30%	28.7/33.5/22.6/10.8/4.4%
GARNET (WHI)	4244	68.35 ± 7.07	100%	161.04 ± 6.01	29.79 ± 5.88	8.50%	22.2/33.9/25.7/12.4/5.8%
GECCO (WHI)	2147	67.30 ± 6.66	100%	161.93 ± 6.10	28.21 ± 5.65	6.00%	32.0/36.2/20.0/7.9/3.9%
HIPFX (WHI)	2995	70.34 ± 6.60	100%	161.42 ± 6.33	26.74 ± 5.31	6.80%	42.6/34.1/16.5/4.7/2.2%
WHIMS+ (WHI)	5666	70.61 ± 6.07	100%	160.65 ± 5.96	28.50 ± 5.46	7.10%	28.7/36.9/22.2/8.5/3.7%
Overall	75232	62.08 ± 12.89	66%	166.72 ± 9.58	28.77 ± 6.10	14.80%	28.6/36.5/21.0/8.7/5.2%

SUPPLEMENTAL ACKNOWLEDGEMENTS

ARIC (phs000280.v3.p1): The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Funding for CARe genotyping was provided by NHLBI Contract N01-HC-65226. Funding for GENEVA was provided by National Human Genome Research Institute grant U01HG004402 (E. Boerwinkle). Individual-participant phenotypic and genotypic data was extracted from the following dbGaP datasets (phs000280.v3.p1, pht004062.v1.p1, pht004063.v1.p1, pht004064.v1.p1, pht004065.v1.p1, pht004032.v1.p1, pht004033.v1.p1, pht004034.v1.p1, pht004035.v1.p1, pht003266.v1.p1, pht000114.v2.p1, phg000079.v1). The authors would like to thank the participants, investigators and staff of the ARIC study for their important contributions.

CARDIA (phs000285.v3.p2): The Coronary Artery Risk Development in Young Adults Study (CARDIA) is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the University of Alabama at Birmingham (N01-HC95095 and N01-HC48047), University of Minnesota (N01-HC48048), Northwestern University (N01-HC48049), and Kaiser Foundation Research Institute (N01-HC48050). This manuscript was not approved by CARDIA. The opinions and conclusions contained in this publication are solely those of the authors, and are not endorsed by CARDIA or the NHLBI and should not be assumed to reflect the opinions or conclusions of either. Funding for CARe genotyping was provided by NHLBI Contract N01-HC-65226. Individual-participant phenotypic and genotypic data was extracted from the following dbGaP datasets (phs000285.v3.p2, pht001583.v2.p2, pht001632.v2.p2, pht001671.v2.p2, pht001720.v2.p2, pht001785.v2.p2, pht001857.v2.p2, pht001589.v2.p2, pht001645.v2.p2, pht001697.v2.p2, pht001744.v2.p2, pht001804.v2.p2, pht001851.v2.p2, pht001557.v2.p2, pht001737.v2.p2, pht001799.v2.p2, pht001845.v2.p2, pht001569.v2.p2, pht001601.v2.p2, pht001656.v2.p2, pht001715.v2.p2, pht001761.v2.p2, pht001811.v2.p2, pht001555.v3.p2, phg000091.v2.p1, phg000092.v2.p1). The authors would like to thank the participants, investigators and staff of the CARDIA study for their important contributions.

CHS (phs000287.v5.p1): The research reported in this article was supported by contract numbers N01-HC-85079, N01-HC-85080, N01-HC-85081, N01-HC-85082, N01-HC-85083, N01-HC-85084, N01-HC-85085, N01-HC-85086, N01-HC-35129, N01 HC-15103, N01 HC-55222, N01-HC-75150, N01-HC-45133, N01-HC-85239 and HHSN268201200036C; grant numbers U01 HL080295 from the National Heart, Lung, and Blood Institute and R01 AG-023629 from the National Institute on Aging, with additional contribution from the National Institute of Neurological Disorders and Stroke. A full list of principal CHS investigators and institutions can be found at http://www.chs-nhlbi.org/pi.html. This manuscript was not prepared in collaboration with

CHS investigators and does not necessarily reflect the opinions or views of CHS, or the NHLBI. Support for the genotyping through the CARe Study was provided by NHLBI Contract N01-HC-65226. Individual-participant phenotypic and genotypic data was extracted from the following dbGaP datasets (phs000287.v5.p1, pht001452.v1.p1, pht001488.v1.p1, pht001489.v1.p1, pht001490.v1.p1, pht001491.v2.p1, pht001492.v1.p1, pht001493.v1.p1, pht001494.v1.p1, pht001495.v1.p1, pht001474.v1.p1, pht001475.v1.p1, phg000077.v1). The authors would like to thank the participants, investigators and staff of the CHS for their important contributions.

Framingham (phs000007.v28.p1): The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195 and HHSN268201500001I). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for CARe genotyping was provided by NHLBI Contract N01-HC-65226. Funding support for the Framingham Metabolomics - Risk Factor Study: Gas Chromatography/Mass Spec - BMI/Lipids/Gluc dataset was provided by FHS contract supplement, NHLBI Intramural Funding. Funding support for the Framingham Central Metabolomics - HILIC - Installment 1 dataset was provided by NIH grant R01DK081572. Funding support for the Framingham Central Metabolomics - HILIC - Installment 2 dataset was provided by NIH grant R01 DK081572. Funding support for the Framingham Metabolomics (HILIC - Installment 1) dataset was provided by NIH grant R01 DK081572. Funding support for the Framingham Metabolomics (HILIC - Installment 2) dataset was provided by NIH grant R01 DK081572. Funding support for the Framingham Metabolomics (Hilic - installment 3) dataset was provided by NIH grant R01 DK081572. Individual-participant phenotypic and genotypic data was extracted from the following dbGaP datasets (phs000007.v28.p1, pht000009.v2.p10, pht000010.v3.p10, pht000011.v3.p10, pht000012.v3.p10, pht000013.v3.p10, pht000014.v3.p10, pht000015.v3.p10, pht000016.v3.p10, pht000017.v3.p10, pht000018.v4.p10, pht000019.v3.p10, pht000020.v3.p10, pht000021.v3.p10, pht000022.v4.p10, pht000023.v4.p10, pht000024.v4.p10, pht000025.v4.p10, pht000026.v3.p10, pht000027.v3.p10, pht000028.v3.p10, pht000029.v3.p10, pht000040.v4.p10, pht003099.v4.p10, pht000030.v7.p10, pht000031.v7.p10, pht000032.v6.p10, pht000033.v8.p10, pht000034.v7.p10, pht000035.v8.p10, pht000036.v8.p10, pht000747.v5.p10, pht000041.v6.p10, pht000074.v9.p10, pht000182.v12.p10, pht001415.v16.p10, pht000183.v12.p10, phg000076.v5). The authors would like to thank the participants, investigators and staff of the Framingham study for their important contributions.

MESA (phs000209.v13.p3): MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-RR-025005, and UL1-TR-000040. The MESA CARe data used for the analyses described in this manuscript were obtained through dbGaP (phs000209.v13.p3,

pht001116.v10.p3, pht001118.v8.p3, pht001119.v8.p3, pht001120.v8.p3, pht003091.v3.p3, phg000081.v2). Funding for CARe genotyping was provided by NHLBI Contract N01-HC-65226. The authors would like to thank the participants, investigators and staff of the MESA study for their important contributions.

COPD (phs000179.v4.p2): This research used data generated by the COPDGene study, which was supported by NIH grants U01HL089856 and U01HL089897. The COPDGene project is also supported by the COPD Foundation through contributions made by an Industry Advisory Board comprised of Pfizer, AstraZeneca, Boehringer Ingelheim, Novartis, and Sunovion. Individual-participant phenotypic and genotypic data was extracted from the following dbGaP datasets (phs000179.v4.p2, pht002239.v3.p2, pht002237.v2.p2, pht002238.v4.p2, phg000490.v1). The authors would like to thank the participants, investigators and staff of the COPD study for their important contributions.

eMERGE (phs000888.v1.p1): ***Group Health Cooperative/University of Washington*** Funding support for Alzheimer's Disease Patient Registry (ADPR) and Adult Changes in Thought (ACT) study was provided by a U01 from the National Institute on Aging (Eric B. Larson, PI, U01AG006781). A gift from the 3M Corporation was used to expand the ACT cohort. DNA aliquots sufficient for GWAS from ADPR Probable AD cases, who had been enrolled in Genetic Differences in Alzheimer's Cases and Controls (Walter Kukull, PI, R01 AG007584) and obtained under that grant, were made available to eMERGE without charge. Funding support for genotyping, which was performed at Johns Hopkins University, was provided by the NIH (U01HG004438). Genome-wide association analyses were supported through a Cooperative Agreement from the National Human Genome Research Institute, U01HG004610 (Eric B. Larson, PI). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000234.v1.p1. ***Mayo Clinic*** samples and associated genotype and phenotype data used in this study were provided by the Mayo Clinic. Funding support for the Mayo Clinic was provided through a cooperative agreement with the National Human Genome Research Institute (NHGRI), Grant #: UOIHG004599; and by grant HL75794 from the National Heart Lung and Blood Institute (NHLBI). Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The datasets used for analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000203.v1.p1. ***Marshfield Clinic Research Foundation*** funding support for the Personalized Medicine Research Project (PMRP) was provided through a cooperative agreement (U01HG004608) with the National Human Genome Research Institute (NHGRI), with additional funding from the National Institute for General Medical Sciences (NIGMS) The samples used for PMRP analyses were obtained with funding from Marshfield Clinic, Health Resources Service Administration Office of Rural Health Policy grant number D1A

RH00025, and Wisconsin Department of Commerce Technology Development Fund contract number TDF FYO10718. Funding support for genotyping, which was performed at Johns Hopkins University, was provided by the NIH (U01HG004438). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000170.v1.p1. ***Northwestern University*** samples and data used in this study were provided by the NUgene Project (www.nugene.org). Funding support for the NUgene Project was provided by the Northwestern University's Center for Genetic Medicine, Northwestern University, and Northwestern Memorial Hospital. Assistance with phenotype harmonization was provided by the eMERGE Coordinating Center (Grant number U01HG04603). This study was funded through the NIH, NHGRI eMERGE Network (U01HG004609). Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000237.v1.p1. ***Vanderbilt University*** funding support for the Vanderbilt Genome-Electronic Records (VGER) project was provided through a cooperative agreement (U01HG004603) with the National Human Genome Research Institute (NHGRI) with additional funding from the National Institute of General Medical Sciences (NIGMS). The dataset and samples used for the VGER analyses were obtained from Vanderbilt University Medical Center's BioVU, which is supported by institutional funding and by the Vanderbilt CTSA grant UL1RR024975 from NCRR/NIH. Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000188.v1.p1. ***Geisinger Health System*** samples and data in this obesity study were provided by the non-alcoholic steatohepatitis (NASH) project. Funding for the NASH project was provided by a grant from the Clinic Research Fund of Geisinger Clinic. Funding support for the genotyping of the NASH cohort was provided by a Geisinger Clinic operating funds and an award from the Clinic Research Fund. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000380.v1.p1. Samples and data in this study were provided by the abdominal aortic aneurysm (AAA) project. Funding for the AAA project was provided by a grant from the Clinic Research Fund of Geisinger Clinic. Funding support for the genotyping of the AAA cohort was provided by a Geisinger Clinic operating funds and an award from the Clinic Research Fund. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession

number phs000387.v1.p1. Samples and data in this study were provided by the Geisinger MyCode® Project. Funding for the MyCode® Project was provided by a grant from Commonwealth of Pennsylvania and the Clinic Research Fund of Geisinger Clinic. Funding support for the genotyping of the MyCode® cohort was provided by a Geisinger Clinic operating funds and an award from the Clinic Research Fund. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000381.v1.p1. ***Mount Sinai School of Medicine*** samples and data used in this study were provided by the Mount Sinai School of Medicine (MSSM) Biobank Project funded by The Charles R. Bronfman Institute for Personalized Medicine (IPM) at Mount Sinai School of Medicine. The Coronary Artery Disease study (IPM BioBank GWAS) is a genome-wide association study funded by the Charles R. Bronfman Institute for Personalized Medicine. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000388.v1.p1. ***The Children's Hospital of Philadelphia (CHOP)*** samples and associated genotype and phenotype data used in this study were provided by the Center for Applied Genomics at the Children's Hospital of Philadelphia. Genotyping for this project was performed at the Center for Applied Genomics and supported by an Institutional DevelopmentAwardfromTheChildren'sHospitalofPhiladelphia. Wegratefullythankallthe children and their families who enrolled in this study, and all individuals who donated blood samples for research purposes. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000490.v1.p1. ***Boston Children's Hospital (BCH)*** samples and data used in this study are provided by The Gene Partnership (TGP) (http://www.genepartnership.org/) a prospective longitudinal study to study the genetic and environmental contributions to childhood health and diseases, collect genetic information on a large number of children who have been phenotyped, and implement the Informed Cohort and the Informed Cohort Oversight Board (ICOB). Children's Hospital Boston (CHB) has committed $10 million for the start-up of the TGP. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000495.v1.p1. ***Cincinnati Children's Hospital Medical Center (CCHMC)*** CCHMC is a participating Pediatric Institution for Phase II of the eMERGE network, a national consortium formed for the purpose of integrating electronic medical records with DNA and sera repositories for large scale, high throughput genetic research. Multiple CCHMC PIs have contributed genome wide association data with various funding support mechanisms. These support mechanisms can be categorized into two groups: disease specific awards (PI initiatives) which focus on particular samples and phenotypes and non-specific awards which contributed to a clinical service. Disease specific awards: 1. Juvenile idiopathic arthritis (JIA): Samples were collected and genotyping was performed by Dr. David Glass with funding support from N01AR42272 and P01AR048929 (PI: Glass). Additional support and genotyping for systemic JIA has been provided by Dr. Dan Kastner's laboratory at the NIH. As of the date of submission, the JIA GWAS data have not been published. 2. Absence seizures: Samples were collected by Dr. Tracy Glauser and genotyping was performed with the support of 5 U01 NS045911 (PI: Glauser) from the National Institute

of Neurological Disorders and Stroke. 3. Autism Spectrum Disorder (ASD): Samples were collected by Drs. Cynthia Molloy and Patricia Manning-Courtney and genotyping was performed with the support of Award 1984, Genome-wide Association Study of Autism Characterized by Developmental Regression (PIs: Molloy & Manning), from Autism Speaks Inc. 4. Eosinophilic Esophagitis: Samples were collected and genotyping was performed by Dr. Marc Rothenberg with funding support of 5 U19 AI066738 Project 3, Eosinophilic esophagitis and food allergy (PI: Sampson, Co-PI & Project 3 PI: Rothenberg). As of the date of the submission, the eosinophilic esophagitis data have not been published. 5. Bicuspid Aortic Valve: Samples were collected and genotyping was performed by Dr. Woodrow Benson with funding support from NIH/NHLBI award HL69712, Genetic mechanisms of cardiac disease in the young (PI: Benson), and NIH/NHLBI award HL74728, SCCOR in Pediatric Heart Development and Disease titled Molecular mechanisms of valve development and disease (PI: Benson).
Non-specific awards: 1. The Cincinnati Control Cohort is a collection of biological samples that have been collected and genotyped through a multidisciplinary approach and with collaboration of more than twenty divisions within CCHMC, supported by the Cincinnati Children's Research Foundation. Lead PIs responsible for this collection are Drs. David Glass and Ardythe Morrow. 2. Clinical cytogenetics samples. Since 2007, more than 2000 samples, enriched for developmental delay, autism and various rare or common genetic diseases as well as specific chromosomal abnormalities such as deletions and duplications, have been genotyped for the purpose of uncovering chromosomal abnormalities. The extraction of data from the EPIC electronic medical record into the de-identified data warehouse, i2b2, was made possible by institutional resources and 1UL1RR026314, Cincinnati Center for Clinical and Translational Sciences and Training Grant (PI: Heubi). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000494.v1.p1. Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number (phs000888.v1.p1,pht004678.v1.p1, pht004677.v1.p1, pht004680.v1.p1, pht005581.v1.p1, pht005587.v1.p1, phg000569.v1, phg000896.v1).

WHI (phs000200.v10.p3): The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C. This manuscript was not prepared in collaboration with investigators of the WHI, has not been reviewed and/or approved by the Women?s Health Initiative (WHI), and does not necessarily reflect the opinions of the WHI investigators or the NHLBI. ***WHI PAGE*** is funded through the NHGRI Population Architecture Using Genomics and Epidemiology (PAGE) network (Grant Number U01 HG004790). Assistance with phenotype harmonization, SNP selection, data cleaning, meta-analyses, data management and dissemination, and general study coordination,

was provided by the PAGE Coordinating Center (U01HG004801-01). ***GARNET*** funding support for WHI GARNET was provided through the NHGRI Genomics and Randomized Trials Network (GARNET) (Grant Number U01 HG005152). Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GARNET Coordinating Center (U01 HG005157). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Funding support for genotyping, which was performed at the Broad Institute of MIT and Harvard, was provided by the NIH Genes, Environment and Health Initiative [GEI] (U01 HG004424). ***WHISP*** the Women?s Health Initiative Sequencing Project (WHISP) was funded by Grant Number RC2 HL102924. This study was part of the NHLBI Grand Opportunity Exome Sequencing Project (GOESP). Funding for GO-ESP was provided by NHLBI grants RC2 HL103010 (HeartGO), RC2 HL102923 (LungGO) and RC2 HL102924 (WHISP). The exome sequencing was performed through NHLBI grants RC2 HL102925 (BroadGO) and RC2 HL102926 (SeattleGO). ***SHARe*** funding for WHI SHARe genotyping was provided by NHLBI Contract N02-HL-64278. ***WHISE*** the WHI Sight Exam and the Memory Study was funded in part by Wyeth Pharmaceuticals, Inc, St. Davids, PA. The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap through dbGaP accession (phs000200.v10.p3, pht000998.v5.p3, pht001019.v5.p3, pht000987.v5.p3, pht000998.v5.p3, phg000592.v1).The authors would like to thank the participants, investigators and staff of the WHI study for their important contributions.