A Fully-Adjusted Two-Stage Procedure for Rank Normalization in Genetic Association Studies: Supplementary Materials

Tamar Sofer, Xiuwen Zheng, Stephanie Gogarten, Cecelia A. Laurie, Kelsey Grinde, John R. Shaffer, Dmitry Shungin, Jeff O'Connell, Ramon A. Durazo-Arvizo, Laura Raffield, Leslie Lange, Solomon Musani, Vasan Ramachandran, Adrienne Cupples, Alex Reiner, Cathy C. Laurie, Kenneth M. Rice

Contents

1 Figure of the various procedures used in the manuscript **3**

2 The Hispanic Community Health Study/Study of Latinos **3**

2.1 Genotyping, imputation, and quality control 4

2.2 Dental exam 4

2.3 HCHS/SOL acknowledgements 4

3 TOPMed **5**

3.1 Genotyping and quality control in TOPMed 5

3.2 TOPMed acknowledgements 6

3.3 The Jackson Heart Study 6

3.4 The Framingham Heart Study 8

3.5 The Amish Study 8

3.6 Comparisons of SKAT p-values between analyses of hemoglobin concentrations 8

4 Comprehensive simulation results **11**

4.1 Explanation of RN-Trait-Adj type 1 error patterns 18

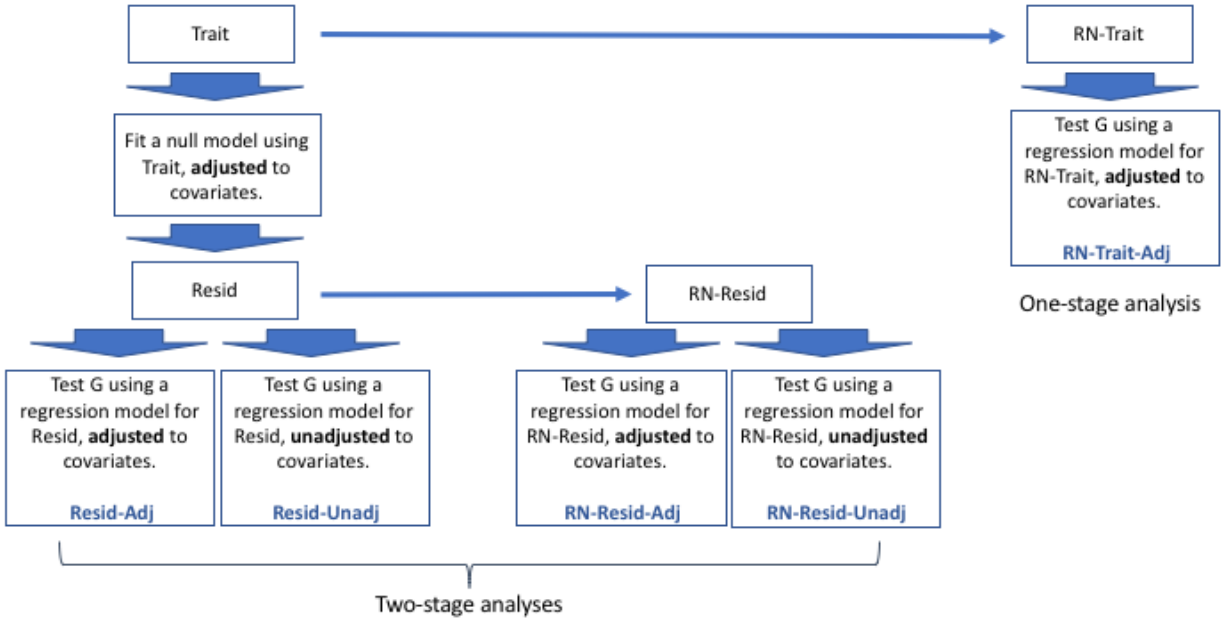
5 Comparison of results from HCHS/SOL GWAS **20**

5.1 NTEETH 21

5.2 BMI 22

5.3 Height 23
5.4 Hip circumference (adjusted to BMI) 24
5.5 Waist circumference (adjusted to BMI) 25
5.6 Waist-to-Hip ratio (adjusted to BMI) 26
5.7 Diastolic blood pressure 27
5.8 Systolic blood pressure 28
5.9 Ferritin 29
5.10 Transferrin 30
5.11 Total iron binding capacity 31
5.12 Hemoglobin A1C 32
5.13 HDL cholesterol 33
5.14 LDL cholesterol 34
5.15 Triglyceride 35
5.16 Total cholesterol 36
5.17 Heart rate 37
5.18 QT interval (electrocardiography) 38
5.19 PR interval (electrocardiography) 39
5.20 Mean corpuscular hemoglobin concentration 40

1 Figure of the various procedures used in the manuscript



2 The Hispanic Community Health Study/Study of Latinos

The HCHS/SOL is a community-based cohort study following 16,415 self-identified Hispanic/Latino participants with initial visits between 2008 and 2011Sorlie et al. (2010). Participants were recruited into the study in four field centers (Chicago, IL, San Diego, CA, Bronx, NY, and Miami, FL) via a two-stage sampling scheme, by which community block units were first sampled, followed by households within the block units. Some or all household members were recruited. The sampling probabilities were set preferentially towards sampling Hispanics/Latinos.LaVange et al. (2010) In total, 12,803 study participants consented to genetic studies. Henceforth, we focus on this subset when describing the HCHS/SOL population. The HCHS/SOL participants are very diverse, and usually self identify as belonging to one of six background groups: Central American, Cuban, Dominican, Mexican, Puerto Rican and South American.

2.1 Genotyping, imputation, and quality control

Blood samples from HCHS/SOL individuals were genotyped the Illumina Omni 2.5M array which was customized to include an additional ∼150,000 markers selected as ancestry-informative markers, variants characteristic of Amerindian populations, known GWAS loci, and candidate gene polymorphisms.

Quality control was similar to the procedure described in Laurie et al. (2010) and included checks for sample identity, batch effects, missing call rate, chromosomal anomalies, deviation from Hardy-Weinberg equilibrium, Mendelian errors, and duplicate sample discordance. A total of 12,803 samples passed quality control, and 2,232,944 SNPs passed quality filters. Pairwise kinship coefficients and principal components reflecting ancestry were estimated using an iterative procedure which accounts for admixture (Conomos et al., 2016). Genome-wide imputation was performed using the 1000 Genomes Project phase 1 reference panel. Genotypes were first pre-phased with SHAPEIT2 (Delaneau et al., 2013) and then imputed with IMPUTE2 (Howie et al., 2009). Each imputed variant was assigned a quality score oevar, defined as the ratio between the observed variance and the expected variance of the allele count.

2.2 Dental exam

Intra-oral examinations were performed following the protocols developed for the National Health and Nutrition Examination Survey (NHANES) (Dye et al., 2007) by trained and calibrated examiners. In brief, the presence or absence of each tooth except third molars (i.e., "wisdom teeth") was recorded. Number of teeth, which takes values of 1 to 28, was defined as the number of teeth (primary, permanent, implant, or root fragment) present at the time of data collection. Dental examinations were not performed for edentulous participants. Figure **??** provides histograms of the residuals of number of teeth ("N-teeth") before and after rank-normalization.

2.3 HCHS/SOL acknowledgements

The authors thank the staff and participants of HCHS/SOL for their important contributions. The Hispanic Community Health Study/Study of Latinos is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina

(HHSN2682013000001I / N01-HC-65233), University of Miami (HHSN268201300004I / N01-HC-65234), Albert Einstein College of Medicine (HHSN268201300002I / N01-HC-65235), University of Illinois at Chicago – HHSN268201300003I / N01-HC-65236 Northwestern Univ), and San Diego State University (HHSN268201300005I / N01-HC-65237). The following Institutes/Centers/Offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements. The HCHS/SOL Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C AM03 and MOD03).

3 TOPMed

3.1 Genotyping and quality control in TOPMed

Whole genome sequencing (WGS) was performed on DNA samples extracted from blood. Sequencing was performed by the Broad Institute of MIT and Harvard (FHS and Amish) and by the Northwest Genome Center (JHS). PCR-free libraries were constructed using commercially available kits from KAPA Biosystems (Broad) or Illumina TruSeq (NWGC). Libraries were pooled for clustering and sequencing, and later de-multiplexed using barcodes. Cluster amplification and sequencing were performed according to manufacturer's protocols using the Illumina cBot and HiSeq X sequencer, to a read depth of >30X. Base calling was performed using Illumina's Real Time Analysis 2 (RTA2) software. Read alignment, variant detection, genotype calling and variant filtering were performed by the TOPMed Informatics Research Center (University of MIchigan). Reads were aligned to the 1000 Genomes hs37d5 decoy reference sequence. Variant detection and genotype calling were performed jointly for several TOPMed studies (including the three analyzed here), using the GotCloud pipeline. Mendelian consistency was used to train a variant quality classifier using a Support Vector Machine, used for variant filtering. Additional quality control (pedigree checks, gender checks, and concordance with prior array data), performed by

the TOPMed Data Coordinating Center, were used to detect and resolve sample identity issues. Further details (including software versions) are provided online (see URL).

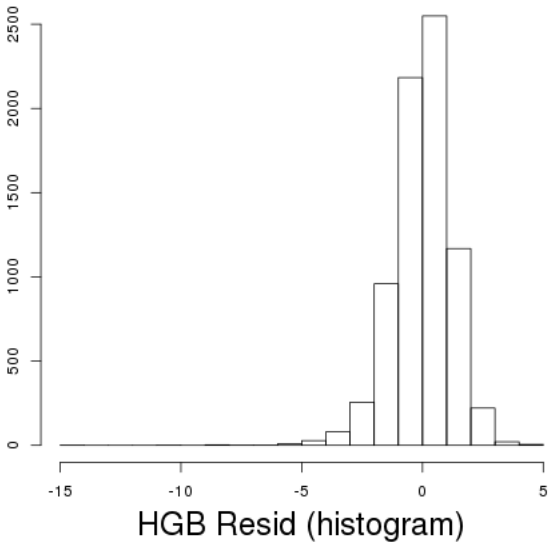
URL: `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/document.cgi?study_id=phs000964.v2.p1&phv=251960&phd=6969&pha=&pht=4838&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1`

3.2 TOPMed acknowledgements

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study" (phs000974.v1.p1) and for "NHLBI TOPMed: Genetics of Cardiometabolic Health in the Old Order Amish Study" (phs000956) were performed at the Broad Institute of MIT and Harvard (HHSN268201500014C). WGS for "NHLBI TOPMed: The Jackson Heart Study" (phs000964.v1.p1) was performed at the University of Washington Northwest Genomics Center (HHSN268201100037C). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). Wegratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

3.3 The Jackson Heart Study

JHS is a longitudinal community -based study designed to assess the causes of the high prevalence of CVD among AAs in the Jackson, Mississippi metropolitan area (Wilson et al., 2005a,b). During the baseline examination period (2000-2004) 5,306 self-identified African Americans were recruited from urban and rural areas of the three counties (Hinds, Madison and Rankin) that comprise the Jackson, Mississippi metropolitan area. Participants were between 35 and 84 years old with the exception of a nested family cohort, where those $\geq$21 years old were eligible All participants included in analyses provided written informed consent for genetic studies. Approval was obtained from the institutional review board of



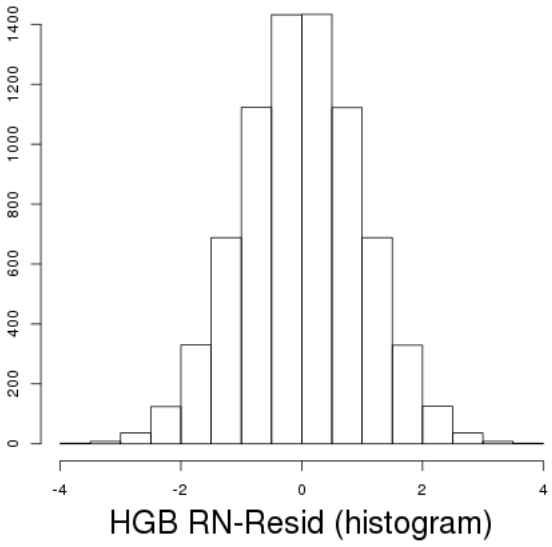


Figure S1: The residual distribution of hemoglobin concentrations in the pooled TOPMed sample (including FHS, JHS, and Amish) from the model adjusted to covariates (top), and after rank-normalization (bottom).

the University of Mississippi Medical Center (UMMC). Data on participants' health behaviors, medical history, and medication use were collected at baseline and subjects underwent venipuncture, allowing for assessment of complete blood cell counts and other measures at UMMC (Beckman-Coulter) (Carpenter et al., 2004).

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201300049C and HHSN268201300050C), Tougaloo College (HHSN268201300048C), and the University of Mississippi Medical Center (HHSN268201300046C and HHSN268201300047C) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS.

3.4 The Framingham Heart Study

The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195 and HHSN268201500001I from the National Heart, Lung and Blood Institute and grant supplement R01 HL092577-06S1 for this research. We also acknowledge the dedication of the FHS study participants without whom this research would not be possible.

3.5 The Amish Study

We gratefully acknowledge our Amish liaisons, research volunteers, field workers and Amish Research Clinic staff and the extraordinary cooperation and support of the Amish community without which these studies would not have been possible. The Amish studies are supported by grants and contracts from the NIH, including U01 HL072515, U01 HL84756, U01 HL137181 and P30 DK72488.

3.6 Comparisons of SKAT p-values between analyses of hemoglobin concentrations

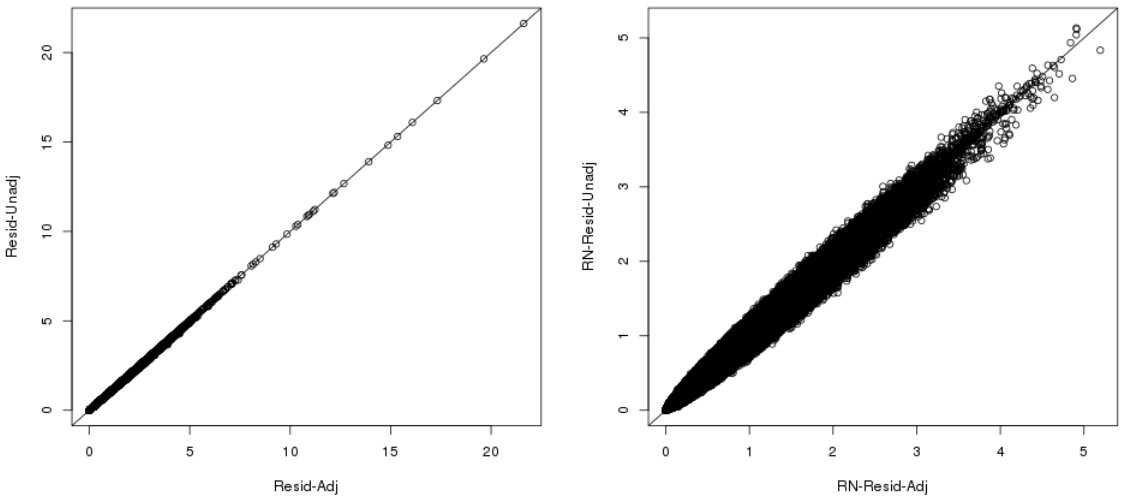


Figure S2: Comparisons of -log(10) of p-values in analyses of HGB in TOPMed, in SKAT applied on variants in non-overlapping windows of length 5Kbp.

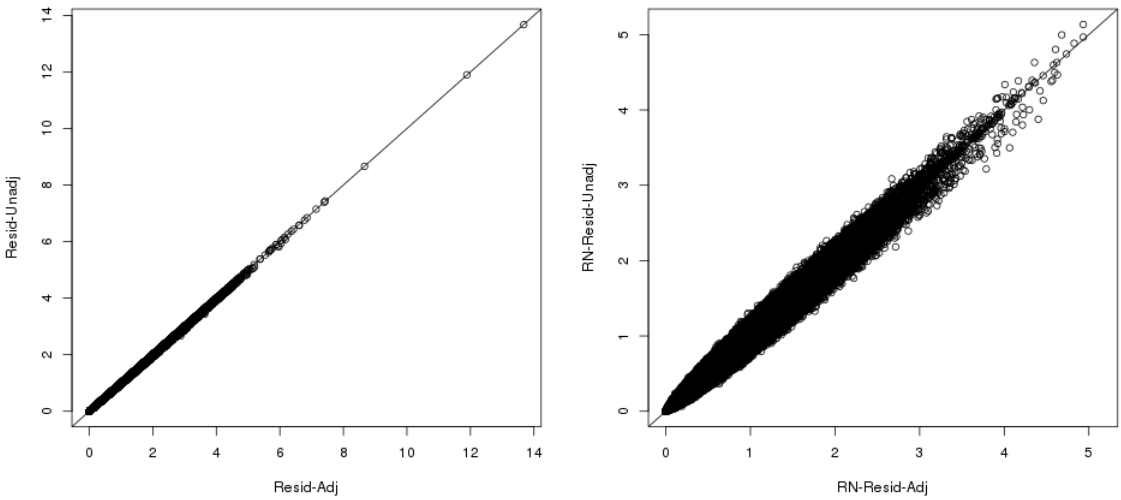


Figure S3: Comparisons of -log(10) of p-values in analyses of HGB in TOPMed, in SKAT applied on variants in non-overlapping windows of length 10Kbp.

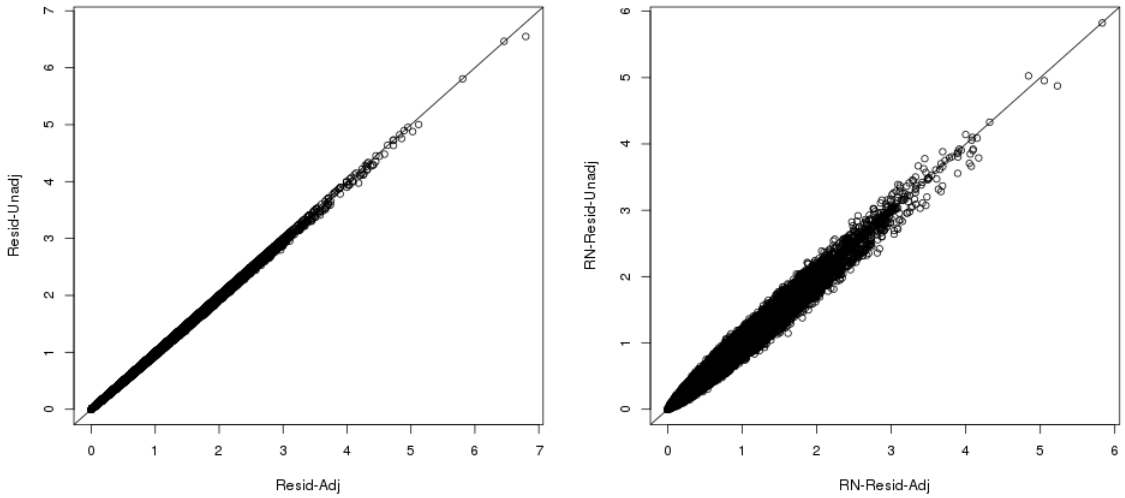
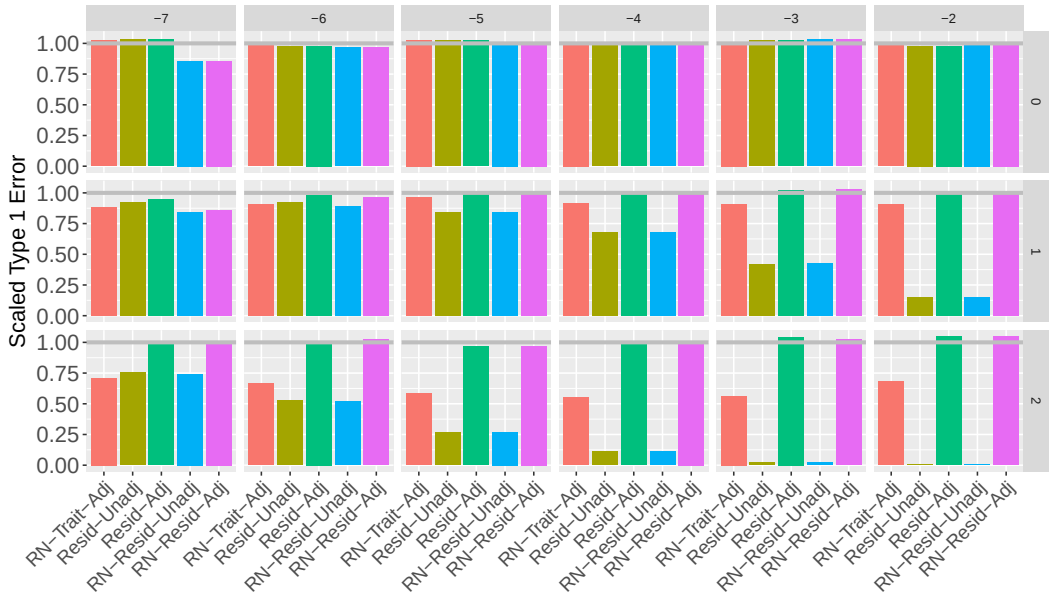


Figure S4: Comparisons of -log(10) of p-values in analyses of HGB in TOPMed, in SKAT applied on variants in non-overlapping windows of length 50Kbp.

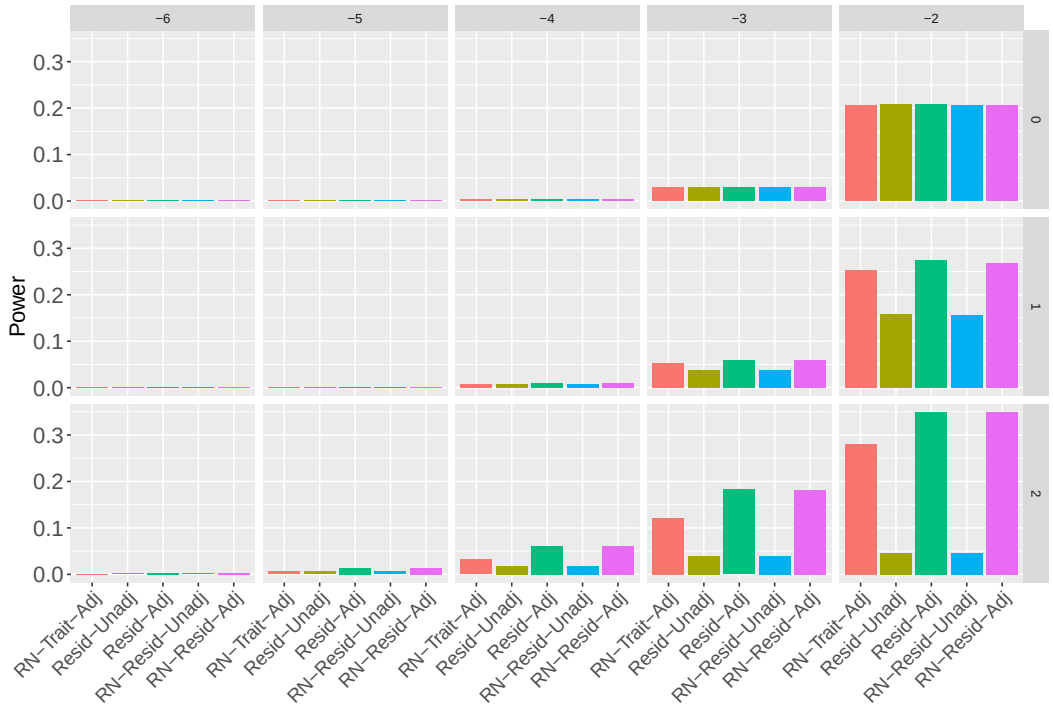
4 Comprehensive simulation results

The following set of figures provide simulation results for each of the outcome settings: "normal", "outlier", and "non-normal", with $n = 2,000$ and $n = 10,000$ sample sizes. For power, we focused on effect sizes $\beta = 0.15$ because patterns were similar for other values, and for both type 1 error and power we present results based on using 10^{-4} as the p-value threshold for declaring significance. Power and type 1 error are calculated as the proportion of simulations in which the null hypothesis was rejected (type 1 error: under the null, with $\beta = 0$). For ease of examination, we scaled the type 1 error by dividing the estimated proportion by the p-value threshold, so that the ideal number if 1. We truncated the scaled type 1 error at 5. In addition, we scaled the power when needed, to match the desired type 1 error rate, as follows: when the type 1 error for a given setting defined by sample size, confounding effect, frequency of genotype, and procedure, was higher than desired (higher than 1×10^{-4}), we performed a binary search to identify the p-value threshold for this settings that results in proportion of false rejections of the null $= 1 \times 10^{-4}$. We used this p-value threshold for calculating power.

In what follows, note that **MAF becomes low as columns are more at the left side of the figures** (intercept is more negative), and **when confounding effect is higher, MAF is slightly increased** (bottom rows in the figure).

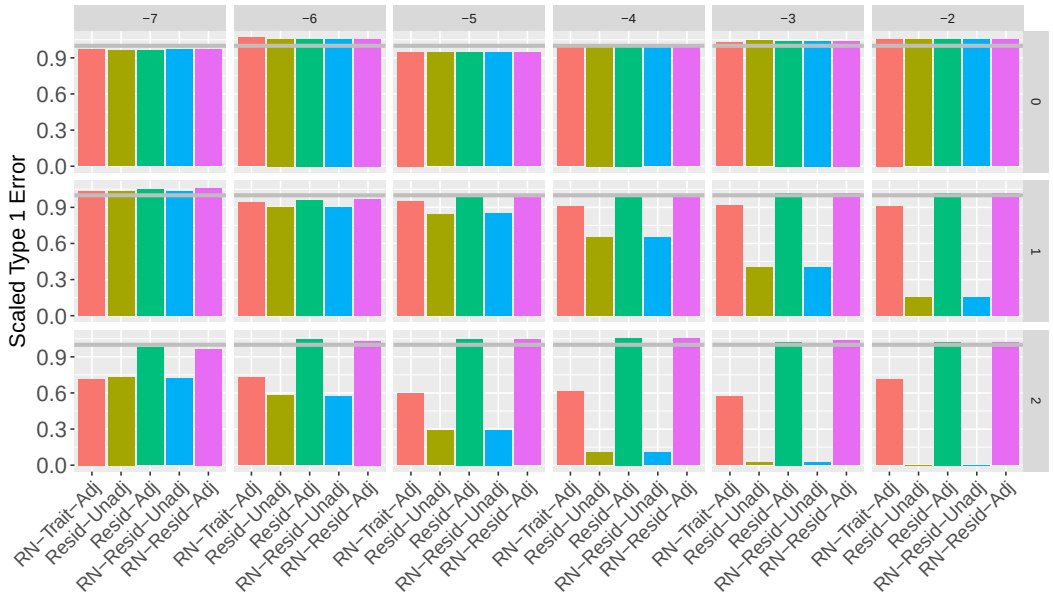


(a) Type 1 error (genotype effet = 0).

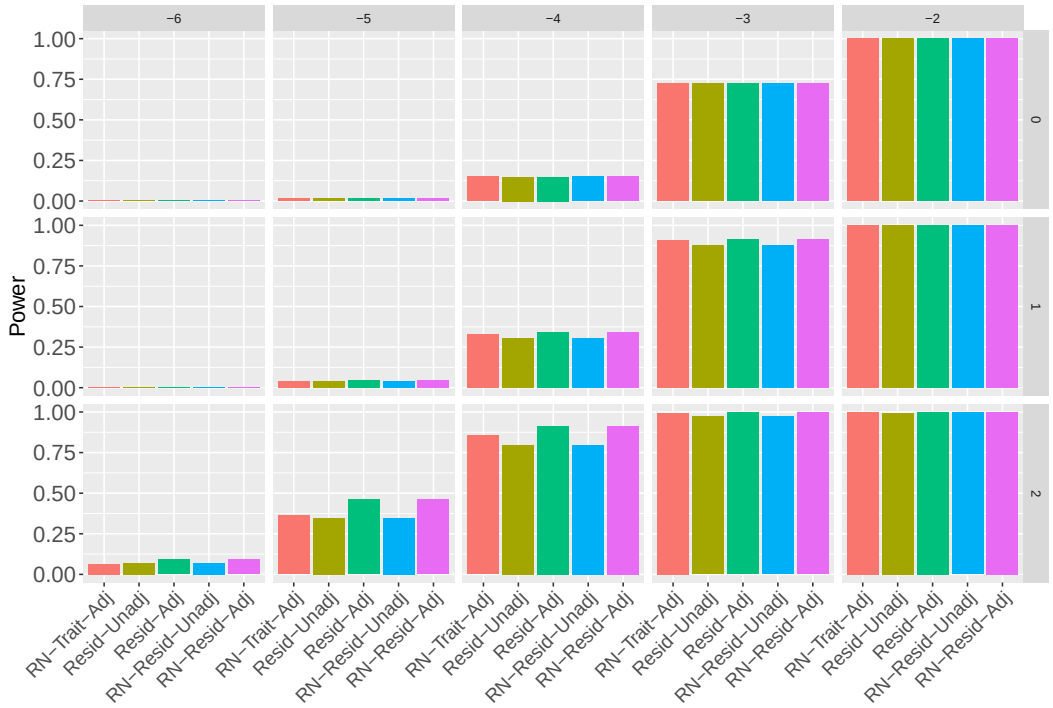


(b) Power (genotype effect = 0.15)

Figure S5: **Normally distributed outcome, n = 2,000.** Type 1 error (top) and power (bottom) from simulations comparing the various analysis one- and two- stage approaches (identified by Outcome_Transformation_Adjustment used in the genotype testing stage) described in the main manuscript.

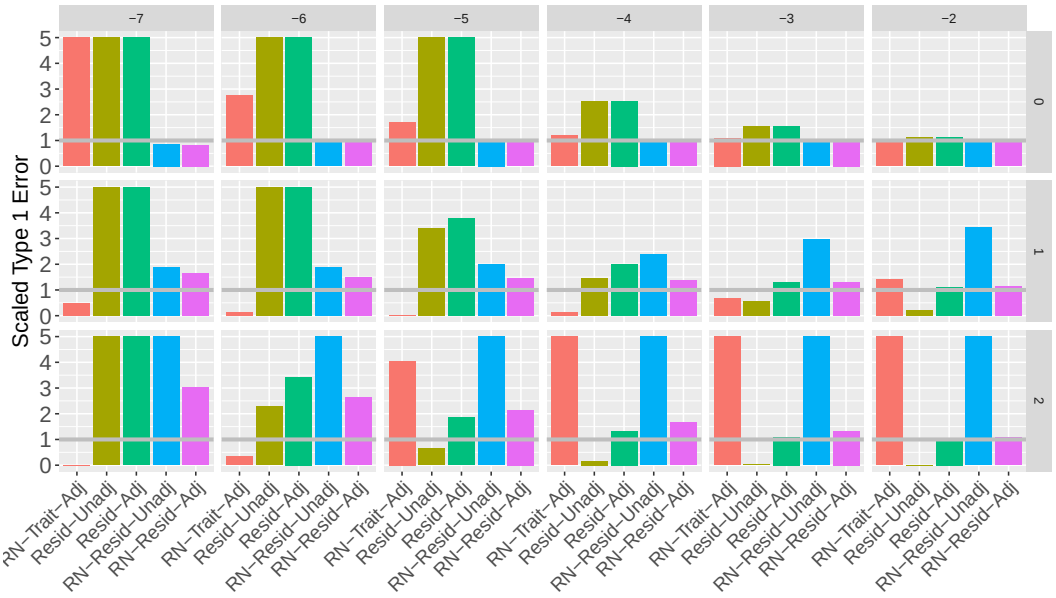


(a) Type 1 error (genotype effet = 0)

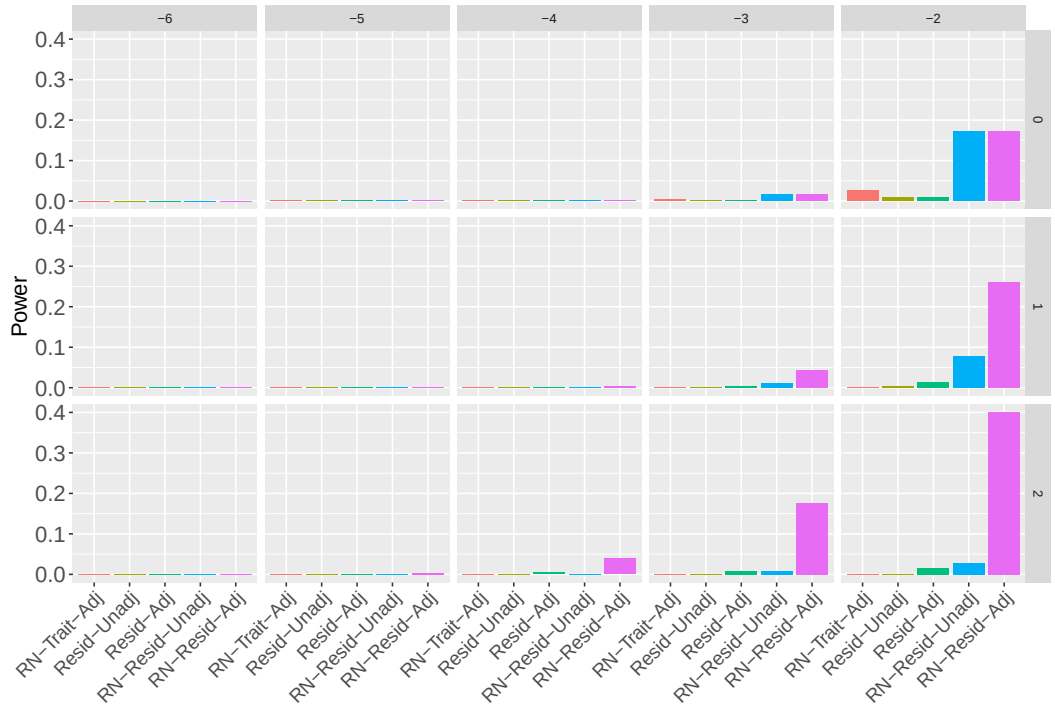


(b) Power (genotype effect = 0.15)

Figure S6: **Normally distributed outcome, n = 10,000.** Type 1 error (top) and power (bottom) from simulations comparing the various analysis one- and two- stage approaches (identified by Outcome_Transformation_Adjustment used in the genotype testing stage) described in the main manuscript.

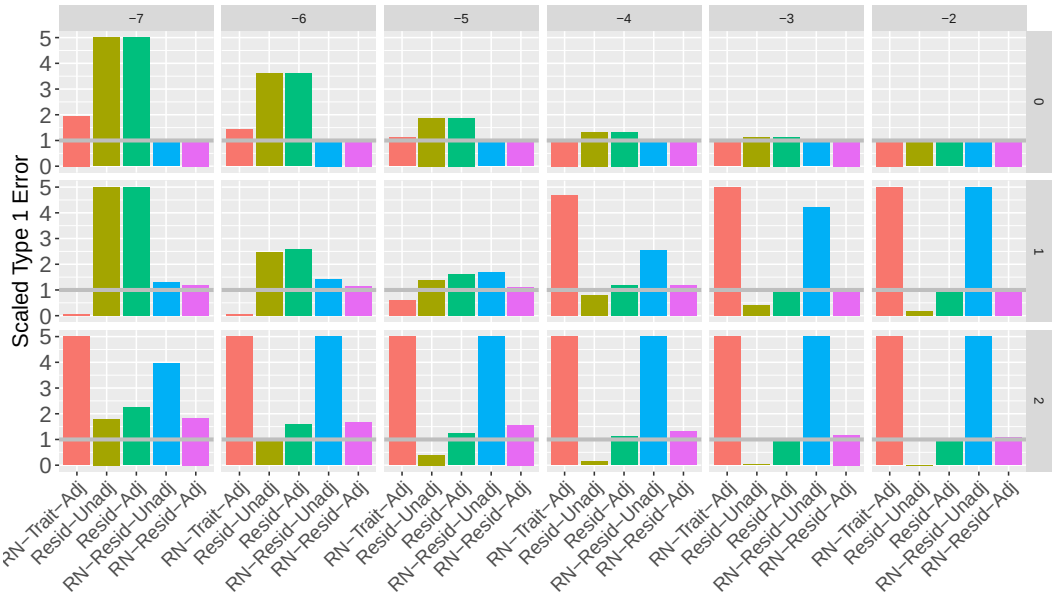


(a) Type 1 error (genotype effet = 0), scaled to 1 and truncated at 5.

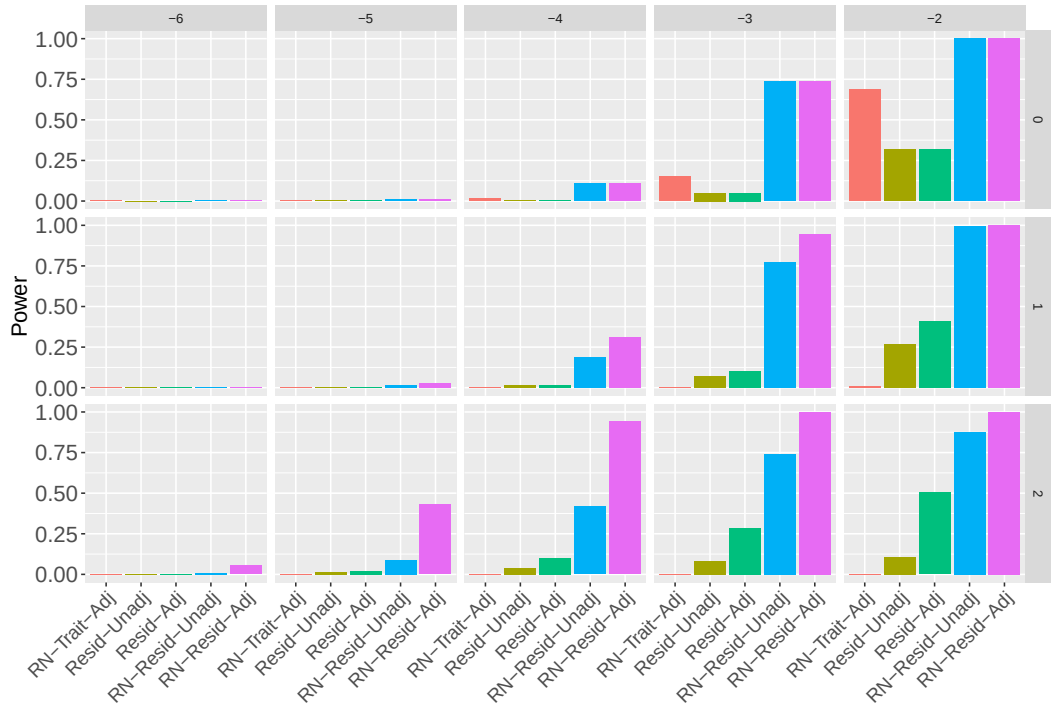


(b) Power (genotype effect = 0.15)

Figure S7: **"non-normal" outcome distribution settings, $\mathbf{n = 2,000}$.** Type 1 error (top) and power (bottom) from simulations comparing the various analysis one- and two- stage approaches (identified by Outcome_Transformation_Adjustment used in the genotype testing stage) described in the main manuscript.

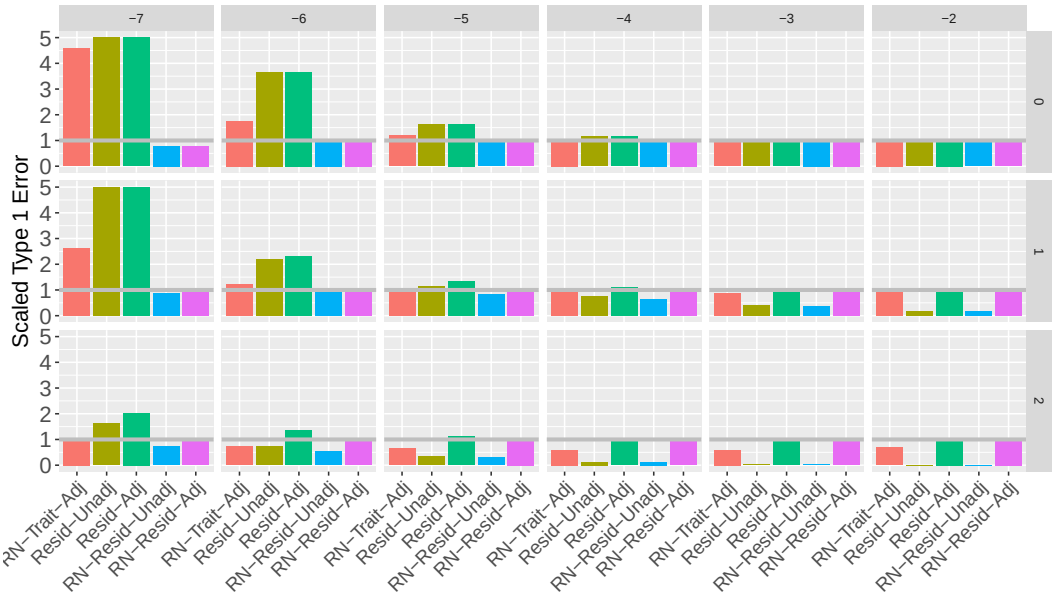


(a) Type 1 error (genotype effet = 0), scaled to 1 and truncated at 5.

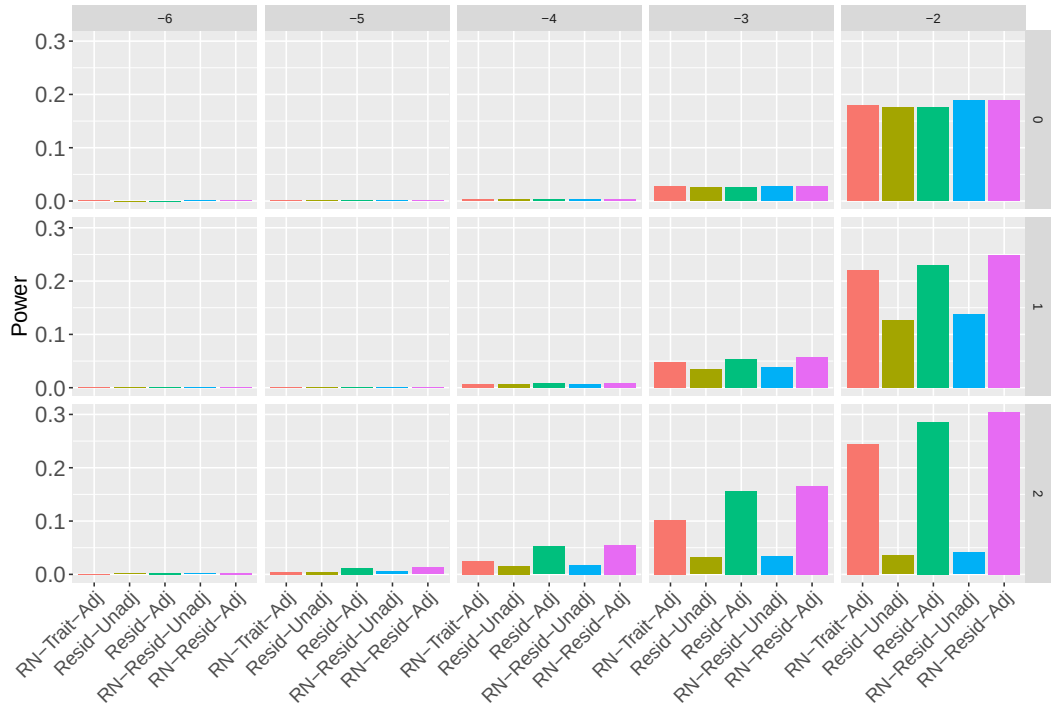


(b) Power (genotype effect = 0.15)

Figure S8: **"non-normal" outcome distribution settings, n = 10,000.** Type 1 error (top) and power (bottom) from simulations comparing the various analysis one- and two- stage approaches (identified by Outcome_Transformation_Adjustment used in the genotype testing stage) described in the main manuscript.

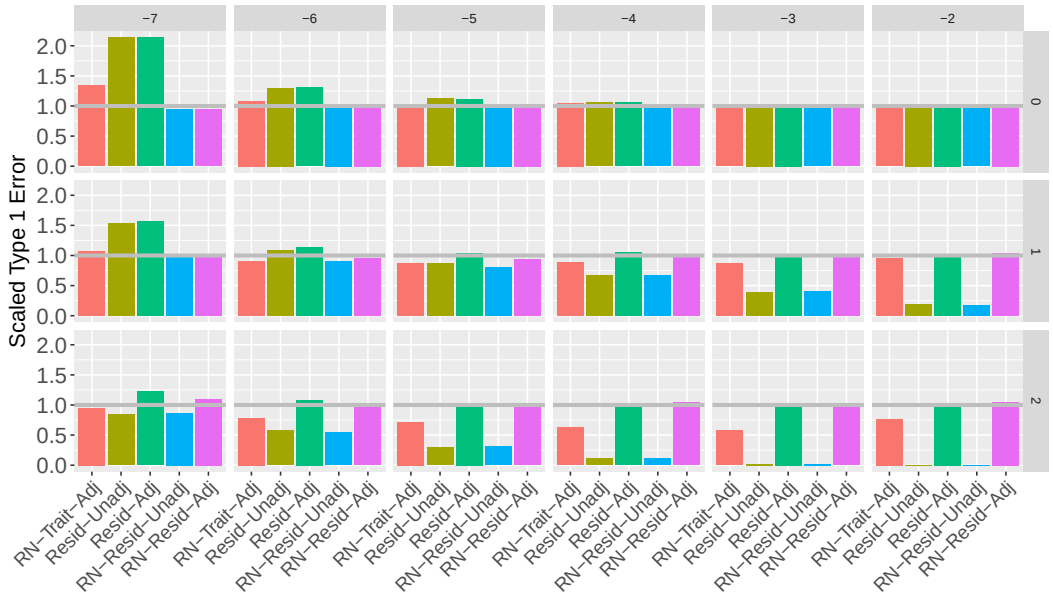


(a) Type 1 error (genotype effet = 0), scaled to 1 and truncated at 5.

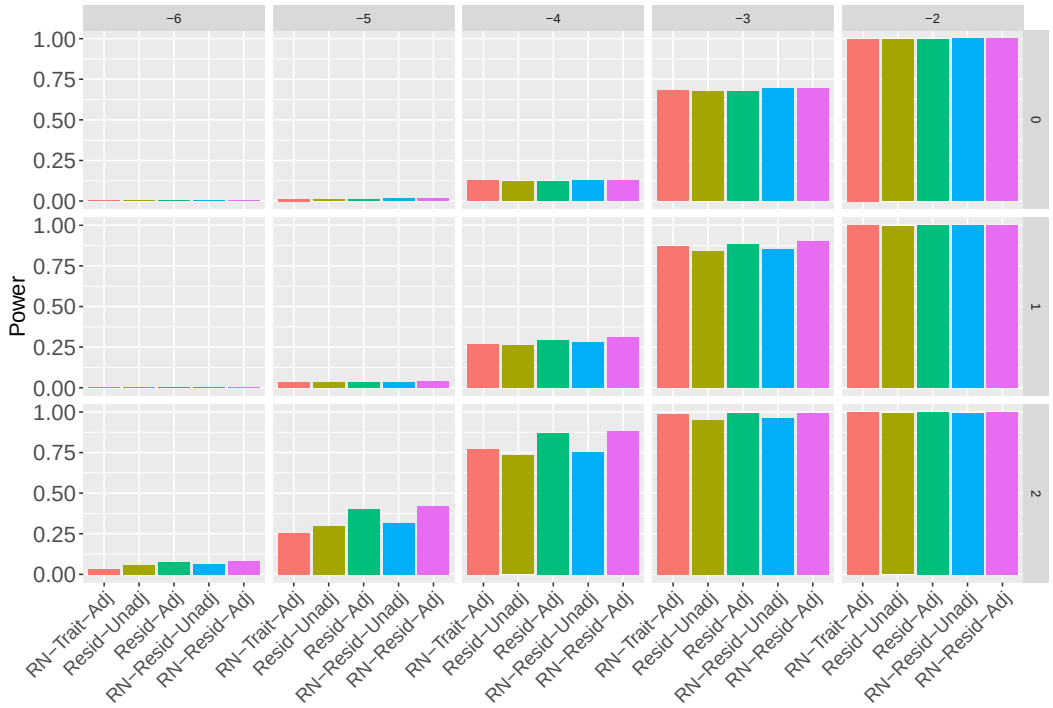


(b) Power (genotype effect = 0.15)

Figure S9: **"Outlier" outcome distribution settings, $\mathbf{n = 2,000}$.** Type 1 error (top) and power (bottom) from simulations comparing the various analysis one- and two- stage approaches (identified by Outcome_Transformation_Adjustment used in the genotype testing stage) described in the main manuscript.



(a) Type 1 error (genotype effet = 0), scaled to 1 and truncated at 5.



(b) Power (genotype effect = 0.15)

Figure S10: **"Outlier" outcome distribution settings, n = 10,000.** Type 1 error (top) and power (bottom) from simulations comparing the various analysis one- and two- stage approaches (identified by Outcome_Transformation_Adjustment used in the genotype testing stage) described in the main manuscript.

4.1 Explanation of RN-Trait-Adj type 1 error patterns

In some settings, and in particular in Figure 2 in the main manuscript, under "How_rare: Rarest", the RN-Trait-Adj analysis demonstrates a surprising pattern, in which the type 1 error is slightly high under no confounding ($\gamma_x = 0$), becomes too conservative under $\gamma_x = 1$, and very high under strong confounding $\gamma_x = 2$. We hypothesized that this could be explained by the mean-variance relationship of TN-trait and the genotype. To investigate this, we perform a small simulation study to obtain mean and variances of RN-trait for levels of the genotype under the confounding settings, and then used the obtained parameter to compute the true versus naïve standard error (SE) of a normal distribution-based test statistics, using the true means and variances (true SE), and the ones combined across levels of the genotype (naïve SE).

Generating data: we generated three data sets, each with 10^{-7} individuals, under the three confounding settings. We adjusted the intercept, so that for $\gamma_x = 0$ we had $\gamma_0 = -6$, for $\gamma_x = 1$ we had $\gamma_0 = -6.5$, and for $\gamma_x = 2$ we had $\gamma_0 = -7.95$. Thus, the three confounding settings resulted in the same allele frequencies. We generated outcomes as described in the manuscript. Table S1 provides the counts of the alleles, as well as means and variances of RN-Trait after regressing the covariate effect (i.e. the residuals) for levels of the genotypes.

Confounding	Genotype level	Number of people	Mean	Variance
0	0	9950700	0.00	0.71
0	1	49,231	0.00	0.70
0	2	69	-0.05	0.68
1	0	9954692	0.00	0.71
1	1	45,151	-0.05	0.50
1	2	157	-0.26	0.32
2	0	9952855	0.00	0.71
2	1	45,926	-0.14	0.38
2	2	1,219	-0.41	0.28

Table S1: Genotype level counts, means, and variances of RN-trait after regressing on the confounder for levels of the genotype and confounding levels. The mean and variance values were used to compute true versus naïve standard error of normal distribution-based test statistics of RN-trait association with the genotype under the null.

The following code provides the function 'inflation' that we used to compute the components of the test statistics, and particularly true versus naïve SEs, and non-centrality parameters of the test statistics.

```
library("sandwich")
```

```
inflate <- function(p, means, vars, n){
   p <- p/sum(p) # the probabilities of genotypes
   g <- 0:2      # genotypes

   gbar <- sum(p*g)
   varg <- sum(p*(g-gbar)^2)

   ybar <- sum(p*means)
   covgy <- sum(p*(means-ybar)*(g-gbar))
   betahat <- covgy/varg
   beta0hat <- ybar - betahat*gbar
   resid <- means - beta0hat-betahat*g

   xmat <- cbind(rep(1, 3), g)
   Bmat <- t(xmat) %*% diag(p) %*% xmat
   Amat <- t(xmat) %*% diag(p) %*% diag( vars + resid^2) %*% xmat

   se.true <- sqrt( (solve(Bmat) %*% Amat %*% solve(Bmat))[2,2]/n )
   se.naiv <- sqrt( solve(Bmat)[2,2] * sum(p*(vars+resid^2))/n )

   return(c(beta0hat=beta0hat, betahat=betahat, se.true=se.true, se.naiv=se.naiv))
}

## we used p-value 0.0001, so:
chithresh= qchisq(1E-4, df=1, lower=FALSE)

inf1 <- inflate(p=c(1E7-49151-157, 49151, 157), means=c(0,-0.05, -0.26),
vars=c(0.71, 0.5, 0.32), n=1E4)
> inf1
     beta0hat       betahat       se.true       se.naiv
 2.492561e-06 -5.101174e-02  1.003762e-01  1.196350e-01
### SE is overestimated, leading to conservative tests
# Proportion of datasets exceeding this threshold:
pchisq( chithresh*(inf1[4]/inf1[3])^2, df=1, ncp=(inf1[2]/inf1[3])^2, lower=FALSE)
 1.836183e-05   ### conservative!

inf2 <- inflate(p=c(1E7-45926-1219, 45926, 1219), means=c(0,-0.14, -0.41),
vars=c(0.71, 0.38, 0.28), n=1E4)
> inf2
     beta0hat       betahat       se.true       se.naiv
 1.439226e-05 -1.462524e-01  8.591739e-02  1.183618e-01
# Proportion of datasets exceeding this threshold:
0.0001273322 ## inflated!
```

5 Comparison of results from HCHS/SOL GWAS

In the following pages, we provide comparisons of results the partly-adjusted (commonly performed) and the fully-adjusted (proposed here) two-stage procedures on various GWAS studies in the HCHS/SOL, to provide some assessment of the degree of loss of power when analyzing relatively normally distributed trait (such as height) without adjustment to covariates at the second stage, and the degree of false positive findings from the same lack of adjustment when analyzing non-normal trait (or more precisely, residuals). All comparisons are provided for common variants (MAF≥ 0.05).

5.1 NTEETH

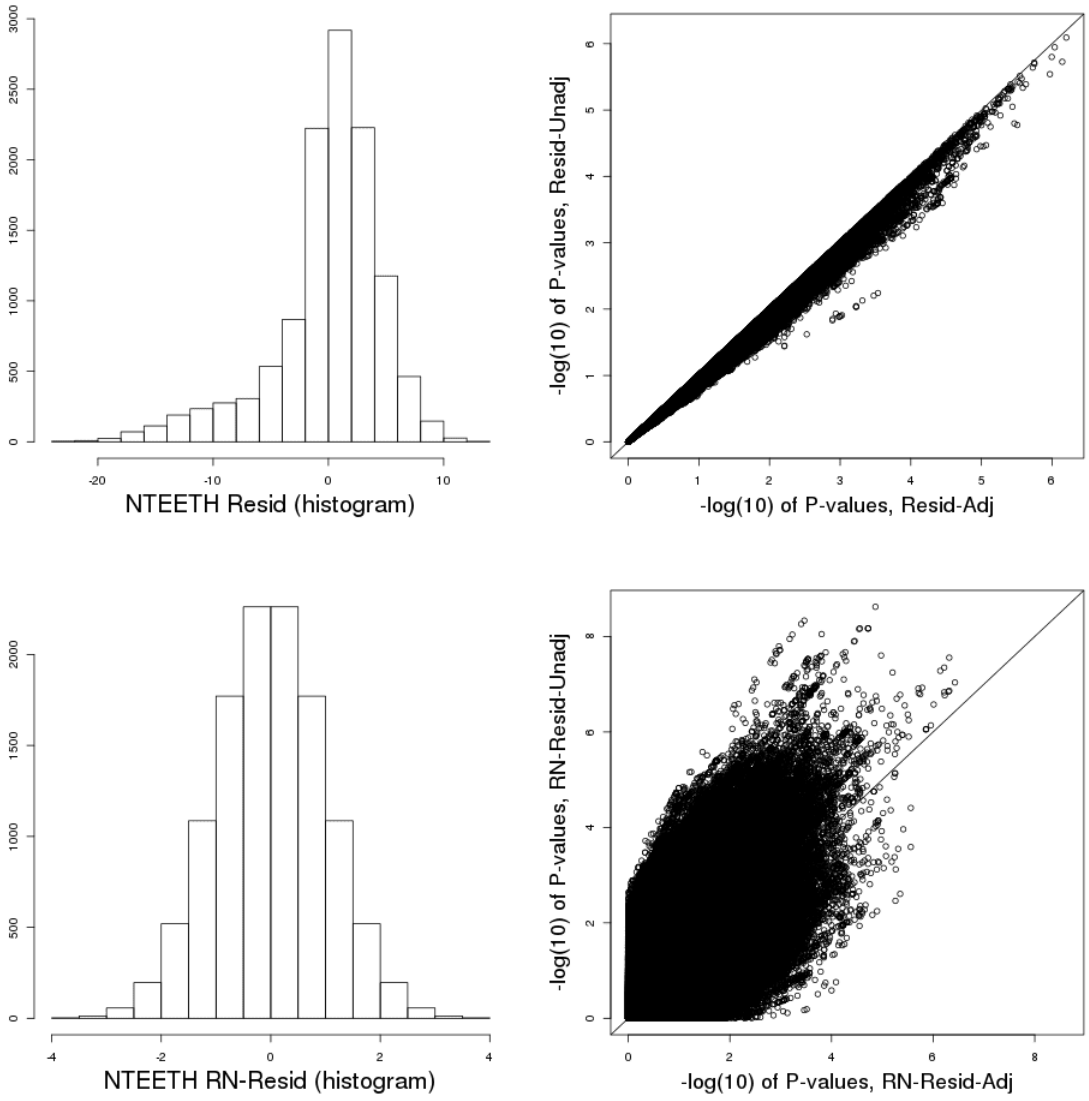


Figure S11: Residuals distribution and GWAS p-values comparisons from the analysis of N-teeth in the HCHS/SOL by various approaches.

5.2 BMI

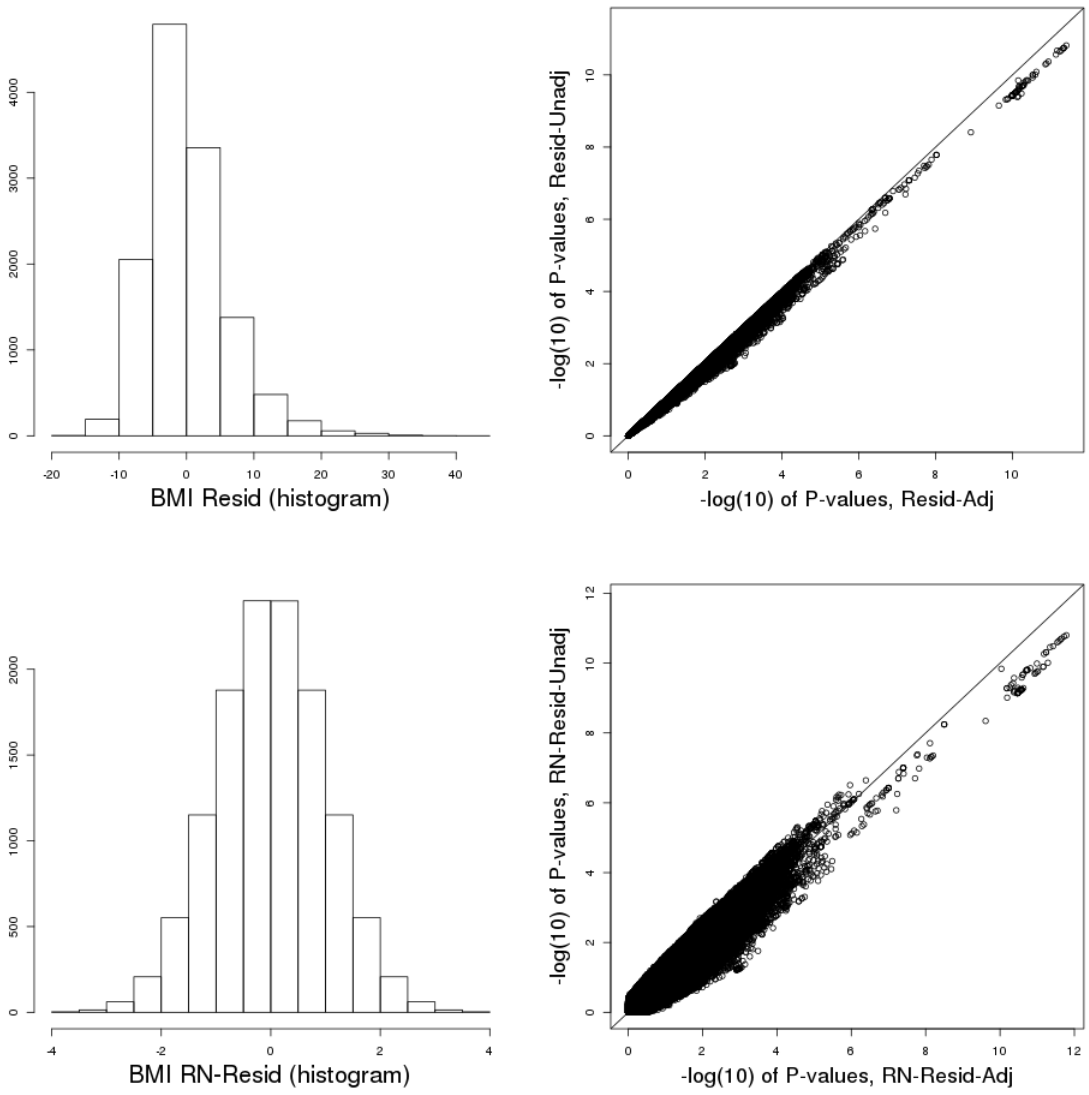


Figure S12: Residuals distribution and GWAS p-values comparisons from the analysis of BMI in the HCHS/SOL by various approaches.

5.3 Height

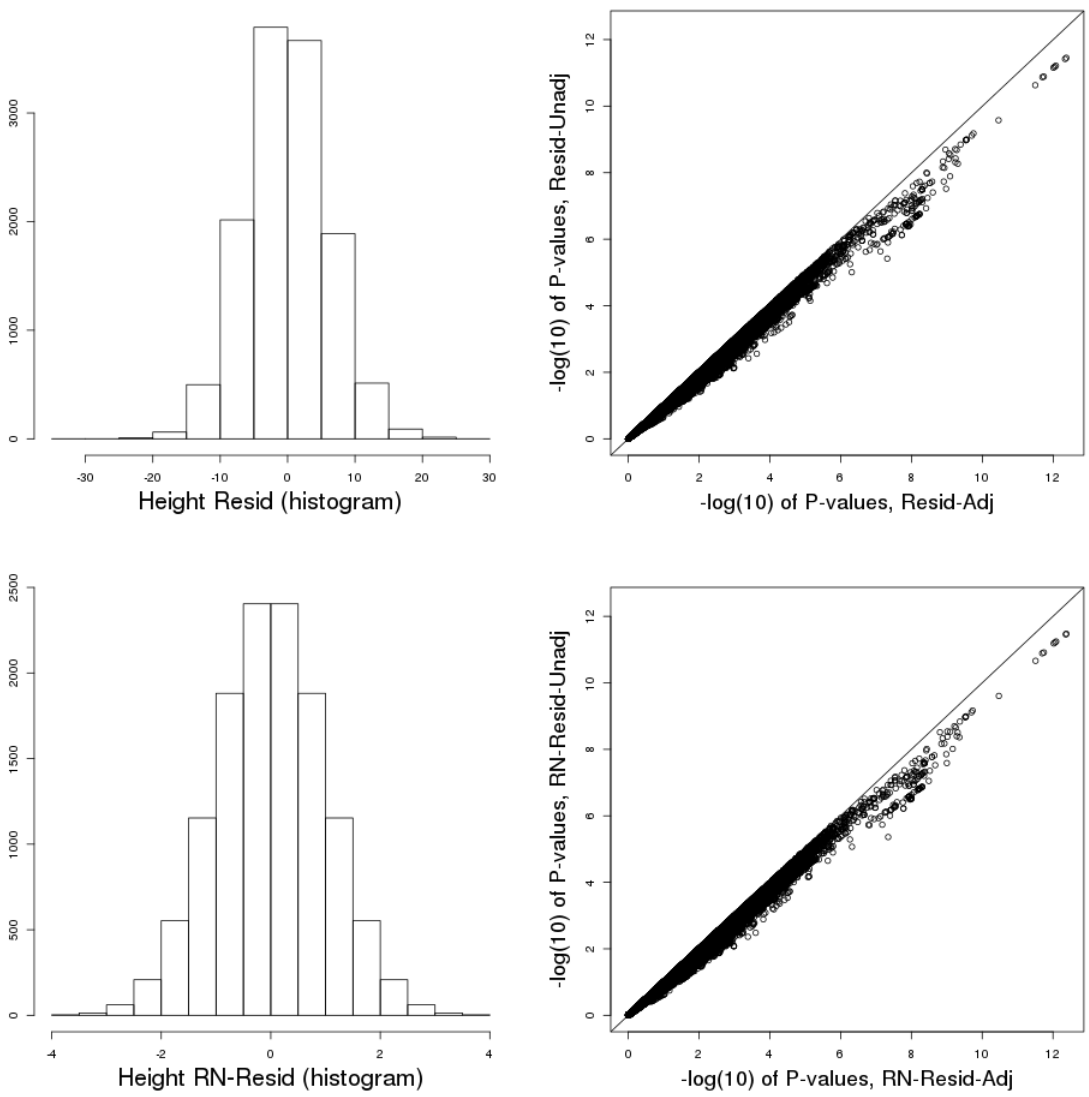


Figure S13: Residuals distribution and GWAS p-values comparisons from the analysis of height in the HCHS/SOL by various approaches.

5.4 Hip circumference (adjusted to BMI)

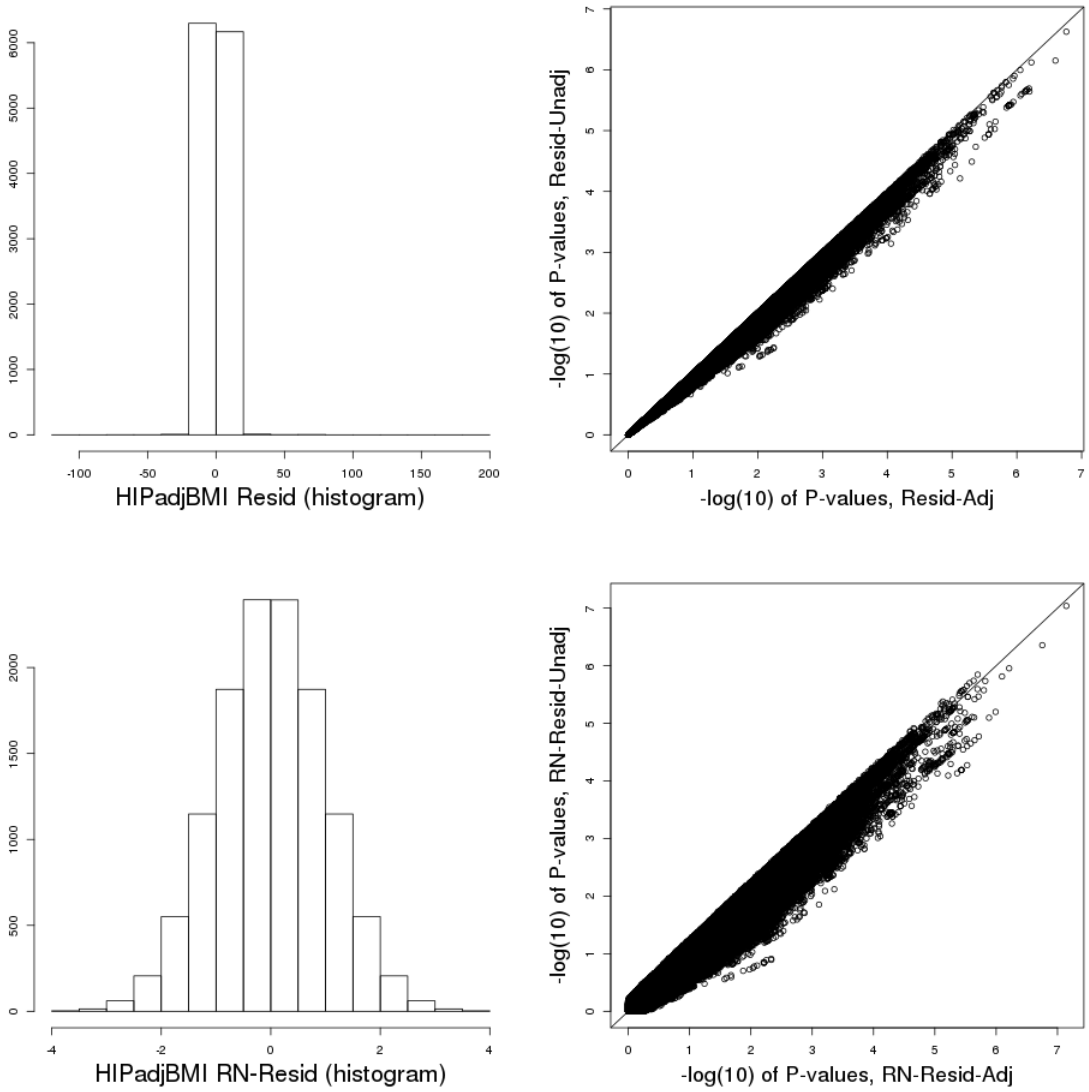


Figure S14: Residuals distribution and GWAS p-values comparisons from the analysis of hip circumference (adjusted to BMI) in the HCHS/SOL by various approaches.

5.5 Waist circumference (adjusted to BMI)

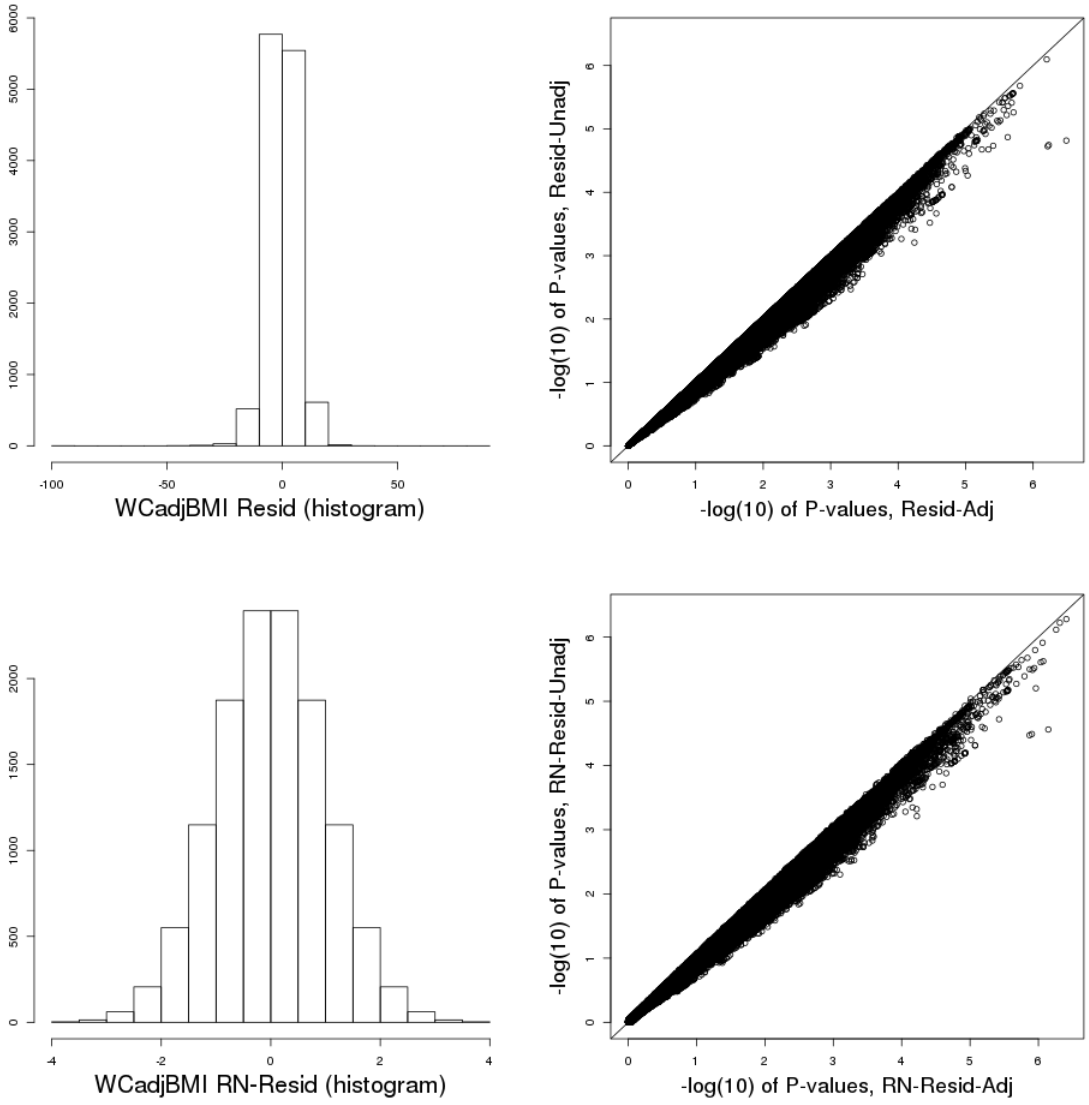


Figure S15: Residuals distribution and GWAS p-values comparisons from the analysis of waist circumference (adjusted to BMI) in the HCHS/SOL by various approaches.

5.6 Waist-to-Hip ratio (adjusted to BMI)

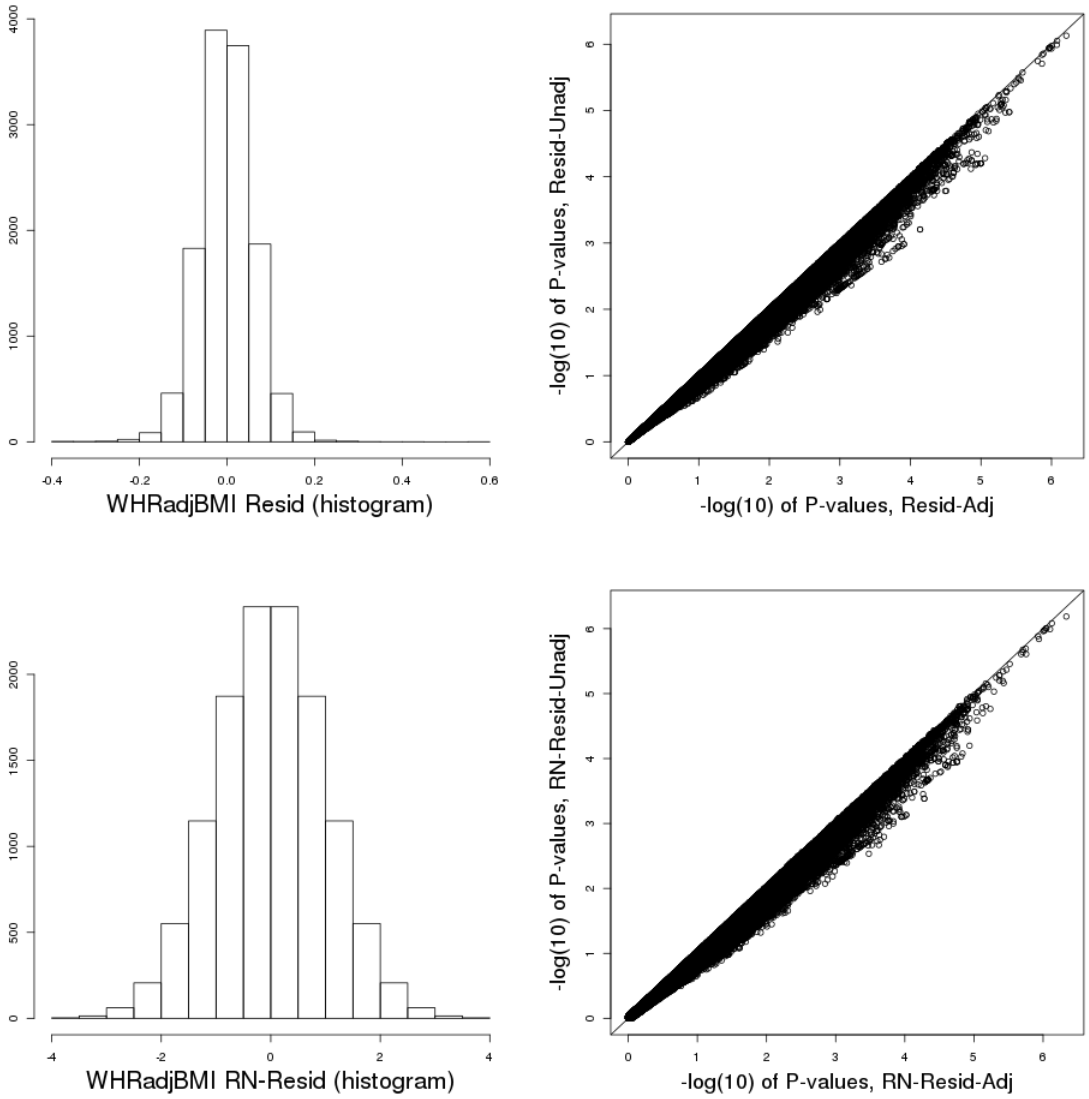


Figure S16: Residuals distribution and GWAS p-values comparisons from the analysis of waist-to-hip ratio (adjusted to BMI) in the HCHS/SOL by various approaches.

5.7 Diastolic blood pressure

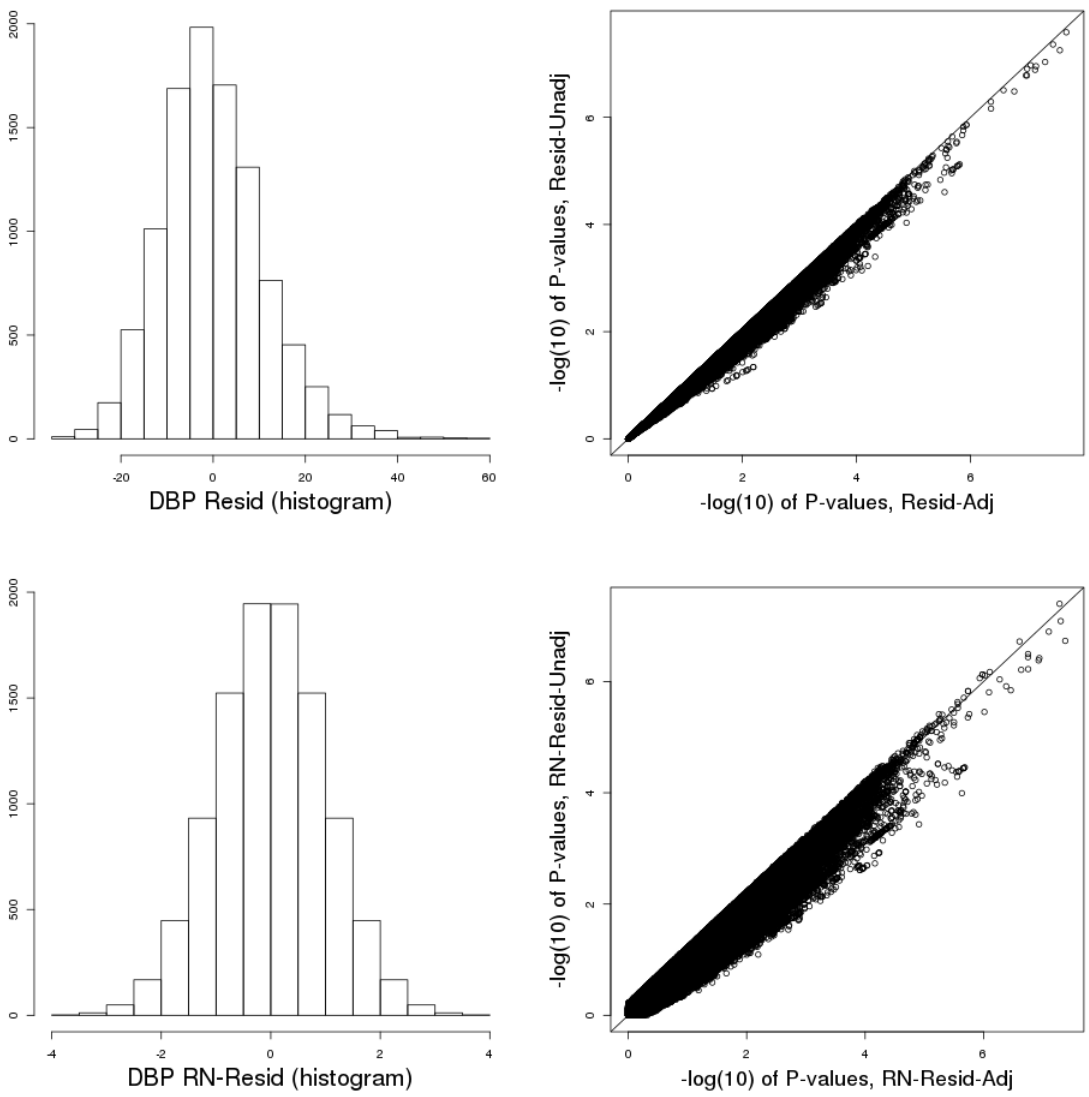


Figure S17: Residuals distribution and GWAS p-values comparisons from the analysis of DBP in the HCHS/SOL by various approaches.

5.8 Systolic blood pressure

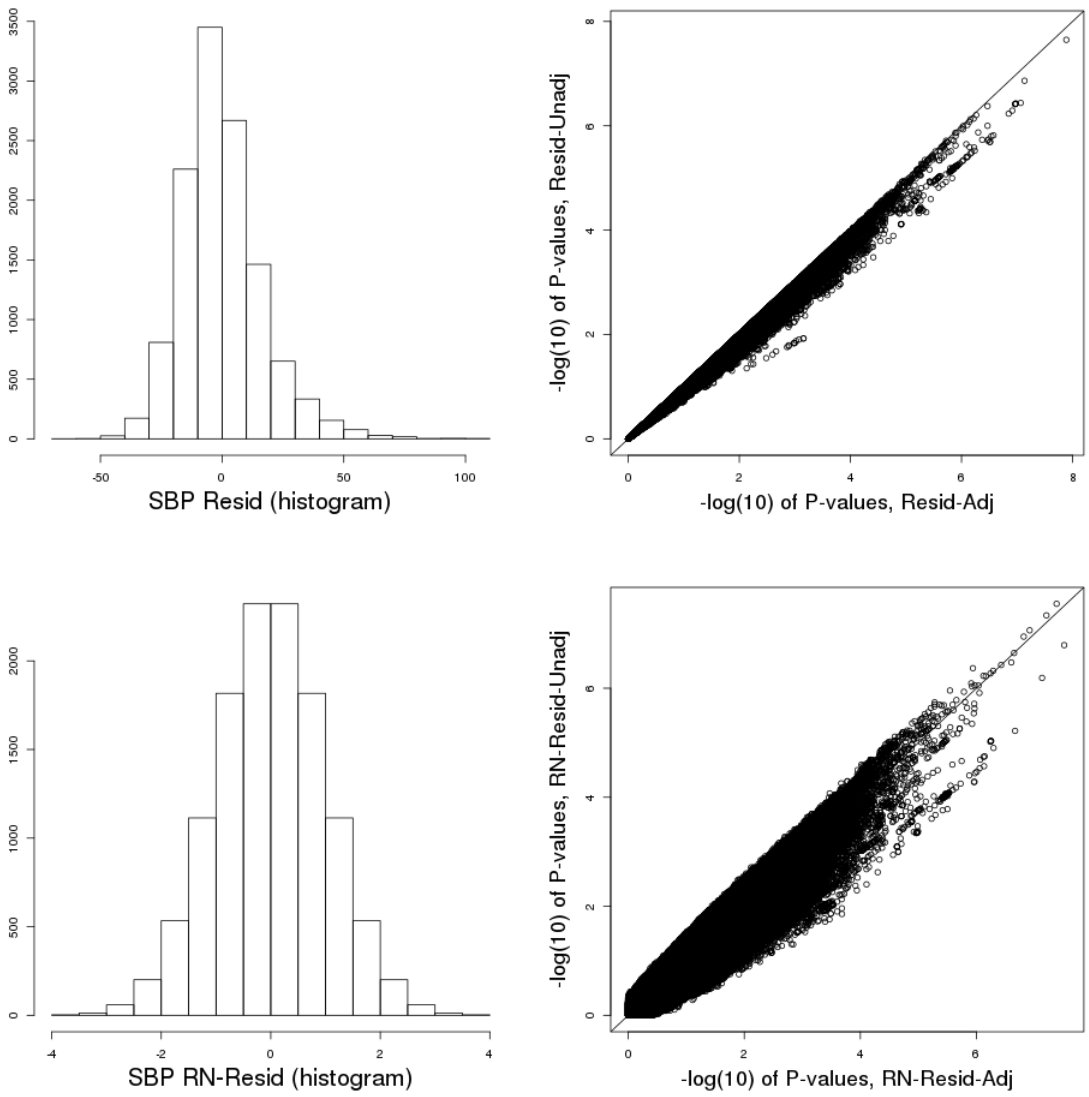


Figure S18: Residuals distribution and GWAS p-values comparisons from the analysis of SBP in the HCHS/SOL by various approaches.

5.9 Ferritin

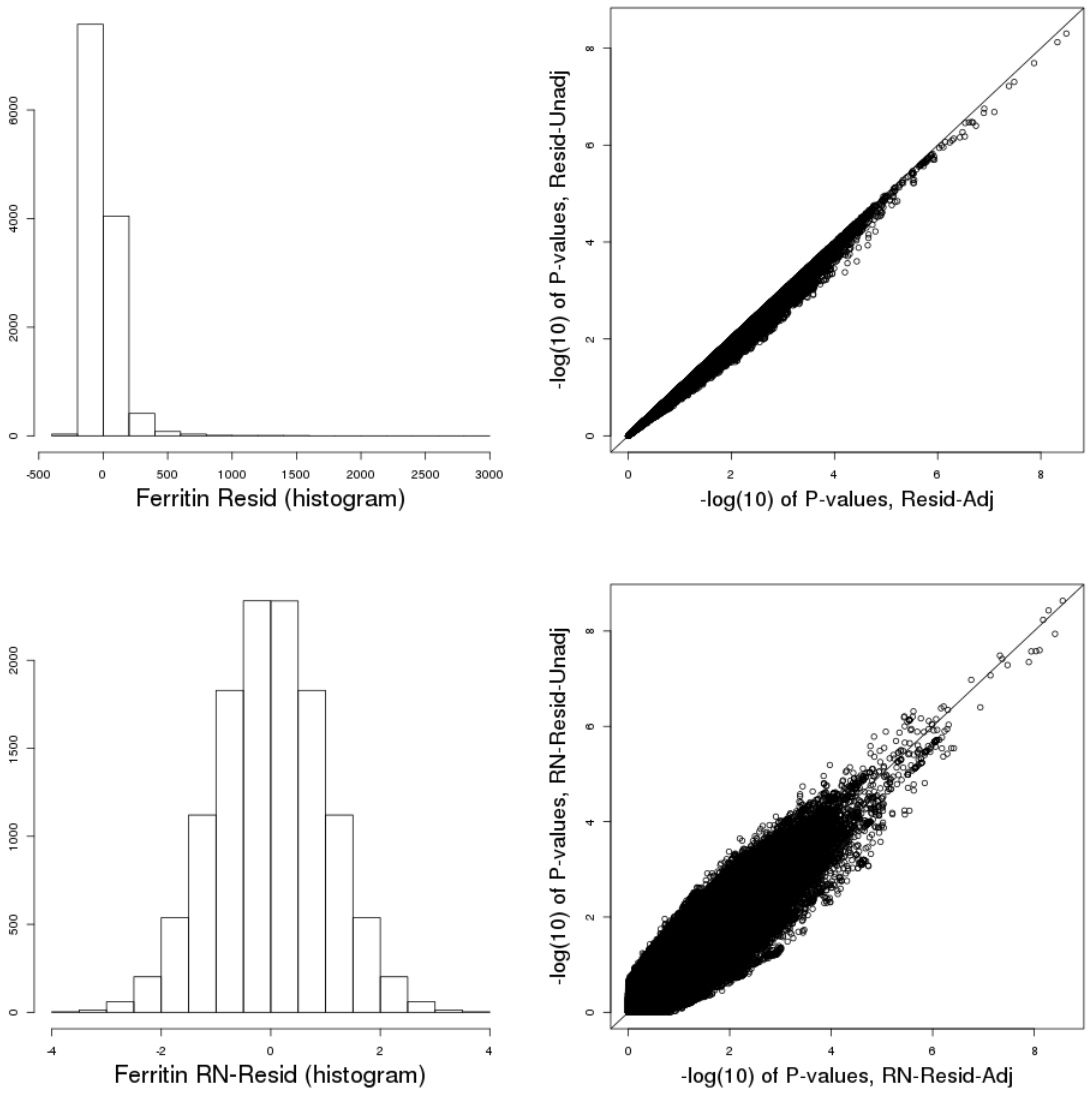


Figure S19: Residuals distribution and GWAS p-values comparisons from the analysis of circulating feritting in the HCHS/SOL by various approaches.

5.10 Transferrin

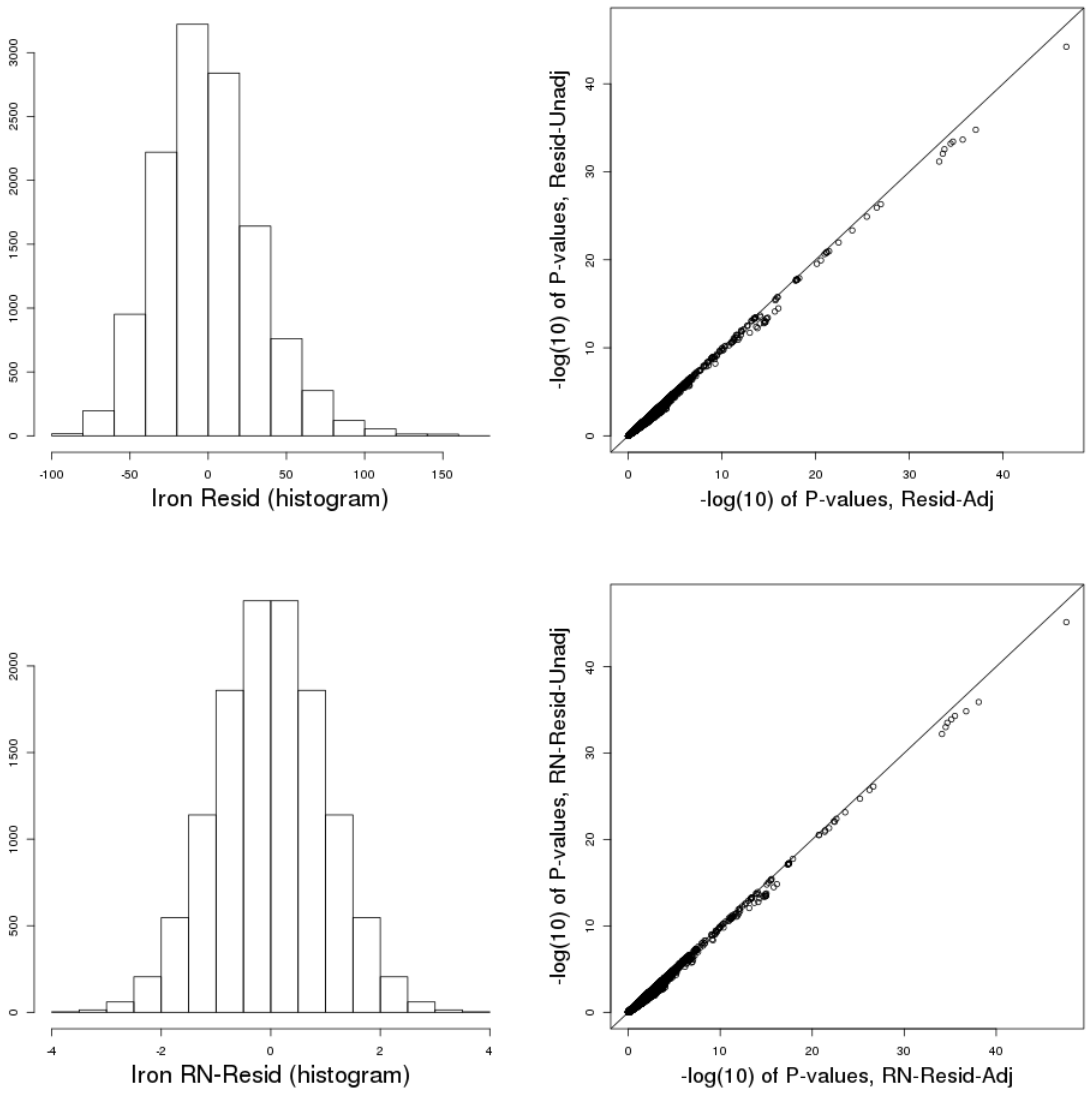


Figure S20: Residuals distribution and GWAS p-values comparisons from the analysis of circulating transferrin in the HCHS/SOL by various approaches.

5.11 Total iron binding capacity

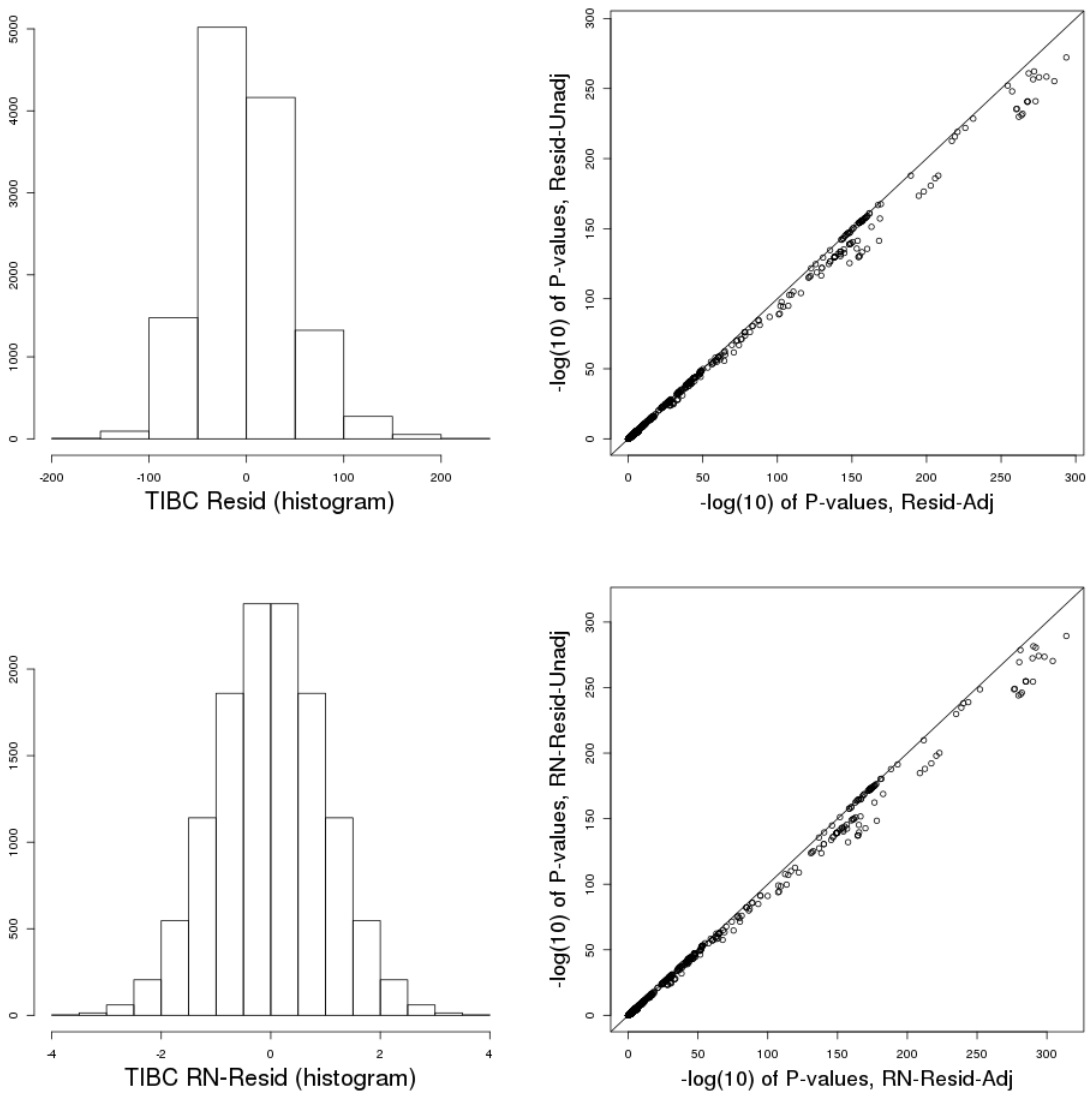


Figure S21: Residuals distribution and GWAS p-values comparisons from the analysis of TIBC in the HCHS/SOL by various approaches.

5.12 Hemoglobin A1C

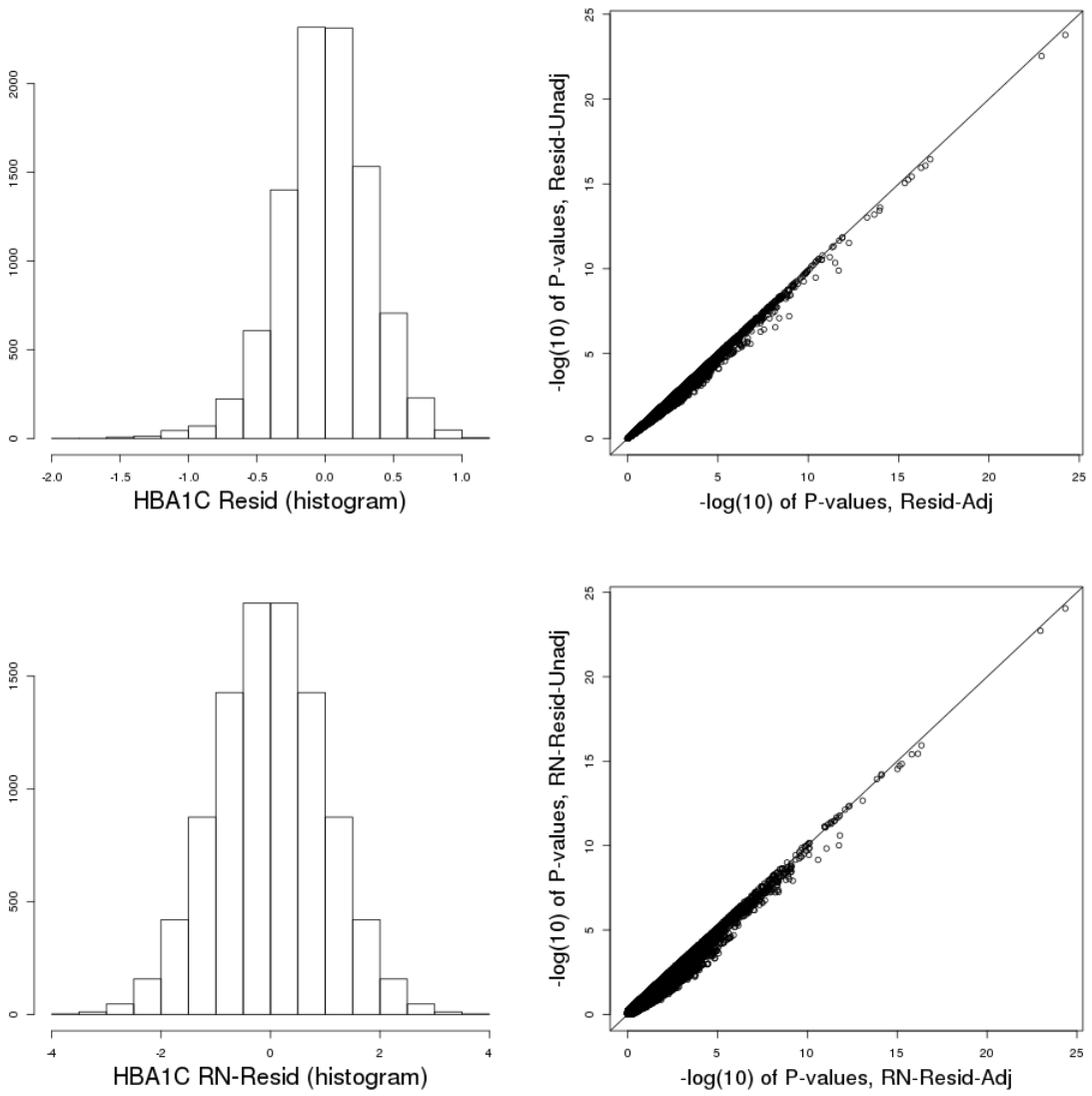


Figure S22: Residuals distribution and GWAS p-values comparisons from the analysis of HBA1C in the HCHS/SOL by various approaches.

5.13 HDL cholesterol

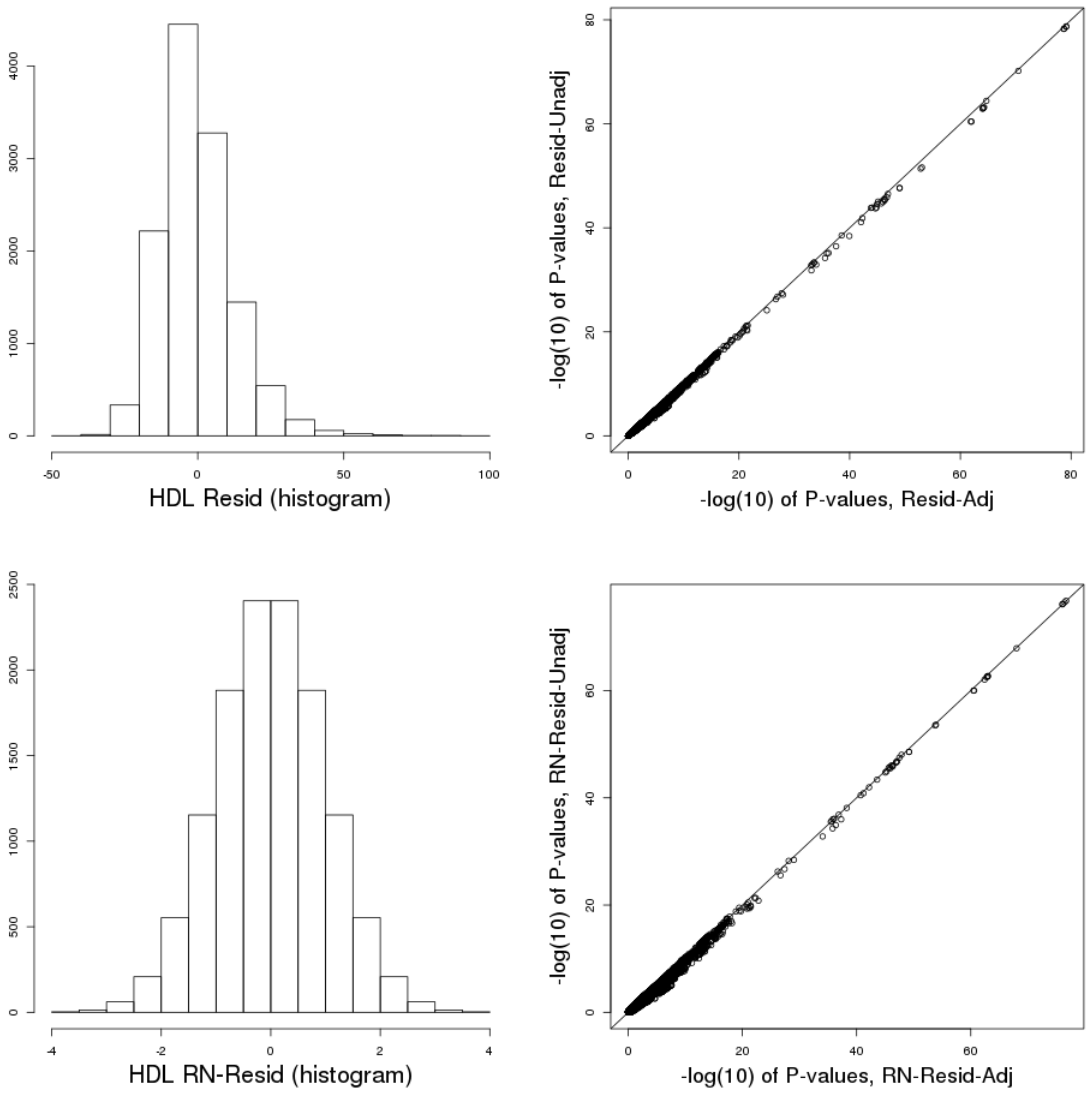


Figure S23: Residuals distribution and GWAS p-values comparisons from the analysis of HDL in the HCHS/SOL by various approaches.

5.14 LDL cholesterol

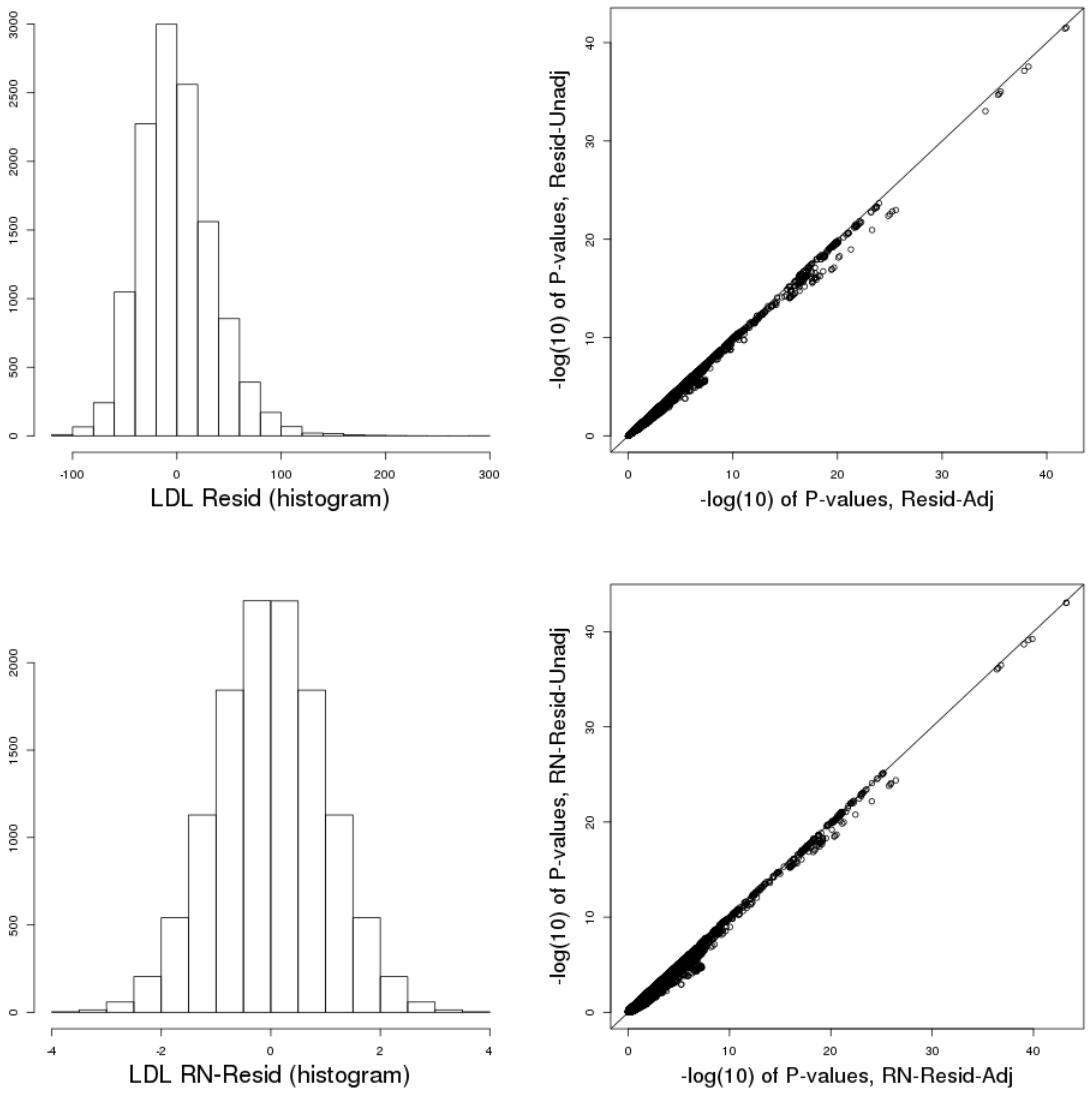


Figure S24: Residuals distribution and GWAS p-values comparisons from the analysis of LDL in the HCHS/SOL by various approaches.

5.15 Triglyceride

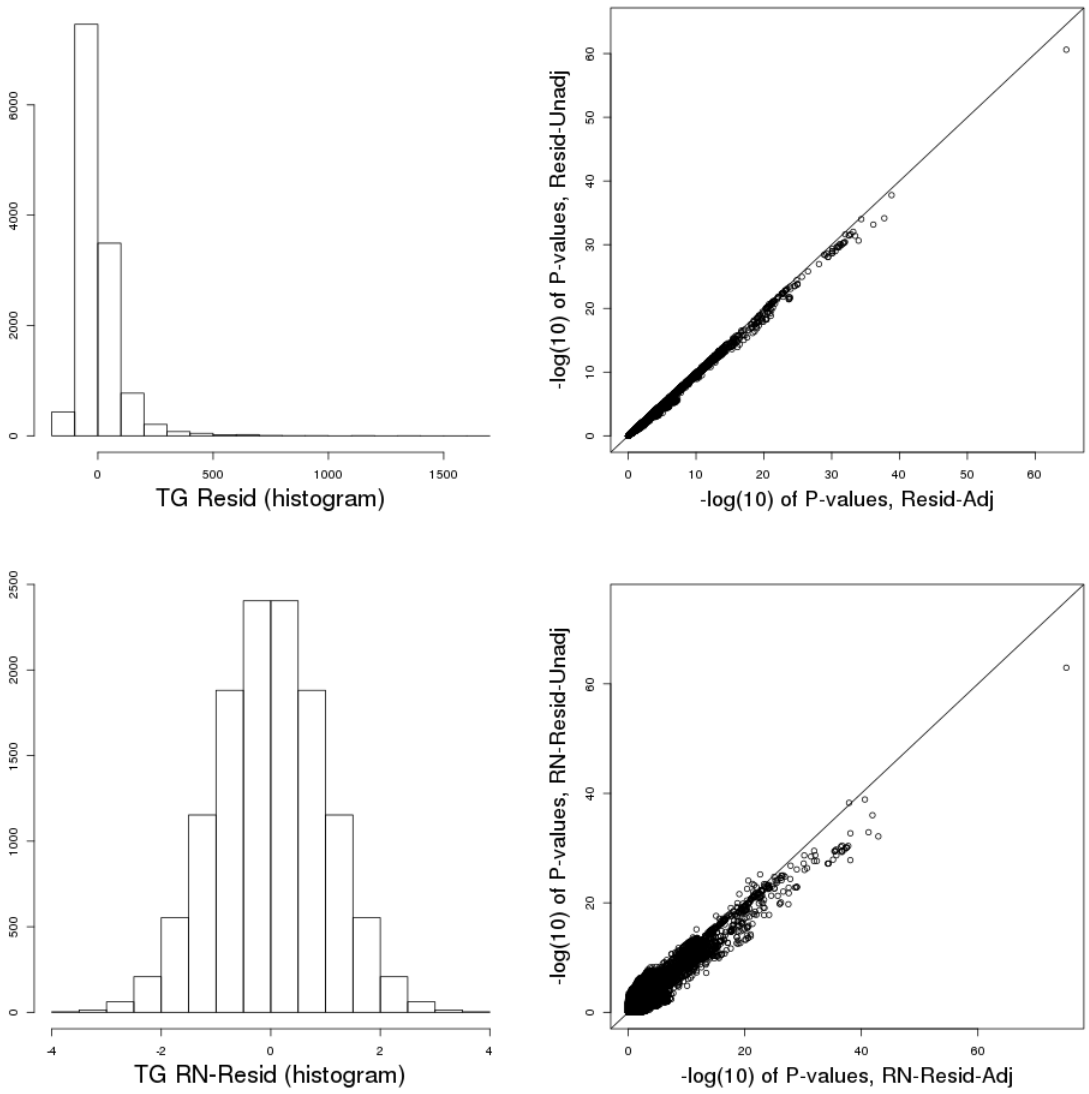


Figure S25: Residuals distribution and GWAS p-values comparisons from the analysis of tryglycerides in the HCHS/SOL by various approaches.

5.16 Total cholesterol

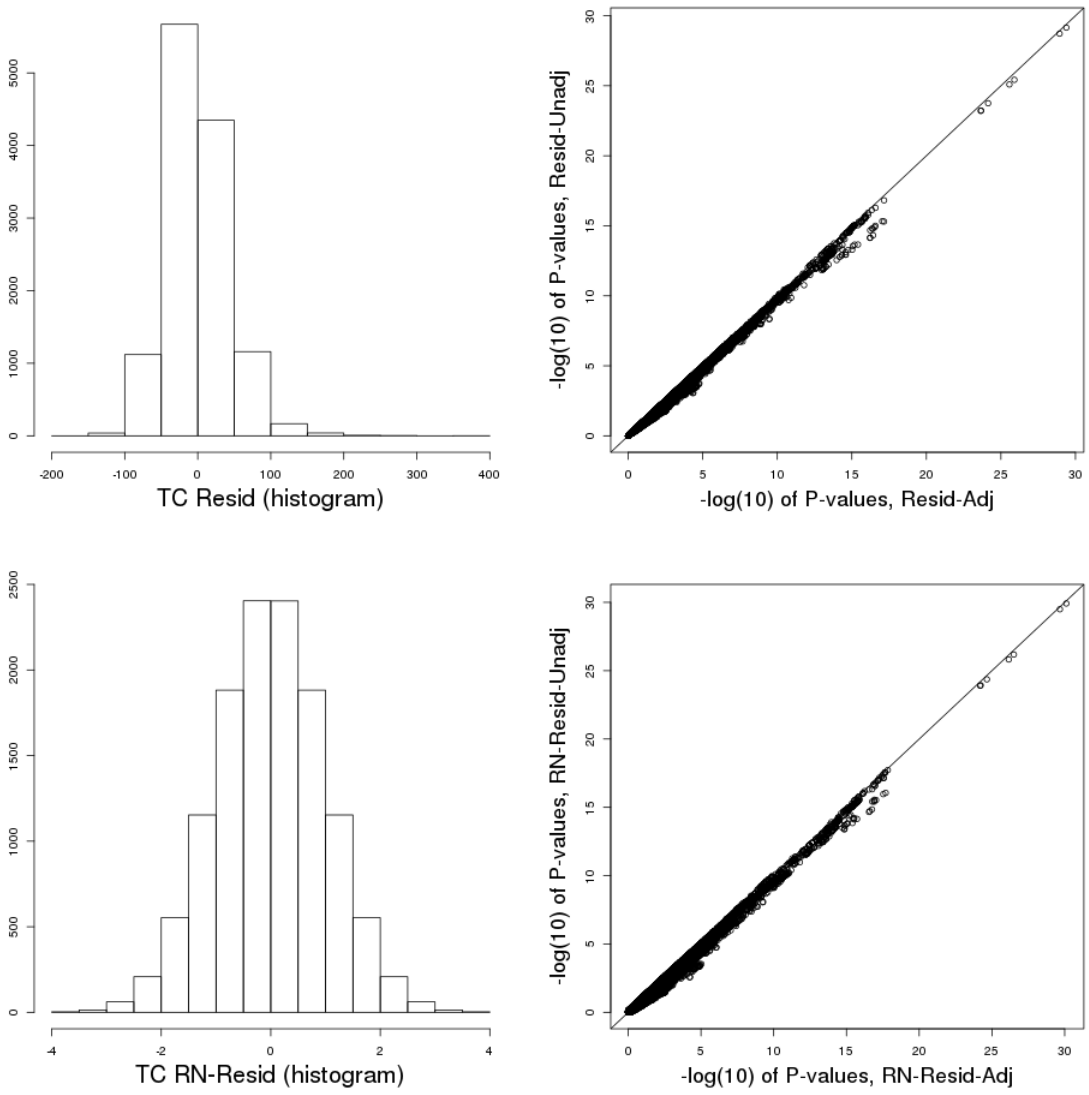


Figure S26: Residuals distribution and GWAS p-values comparisons from the analysis of total cholesterol in the HCHS/SOL by various approaches.

5.17 Heart rate

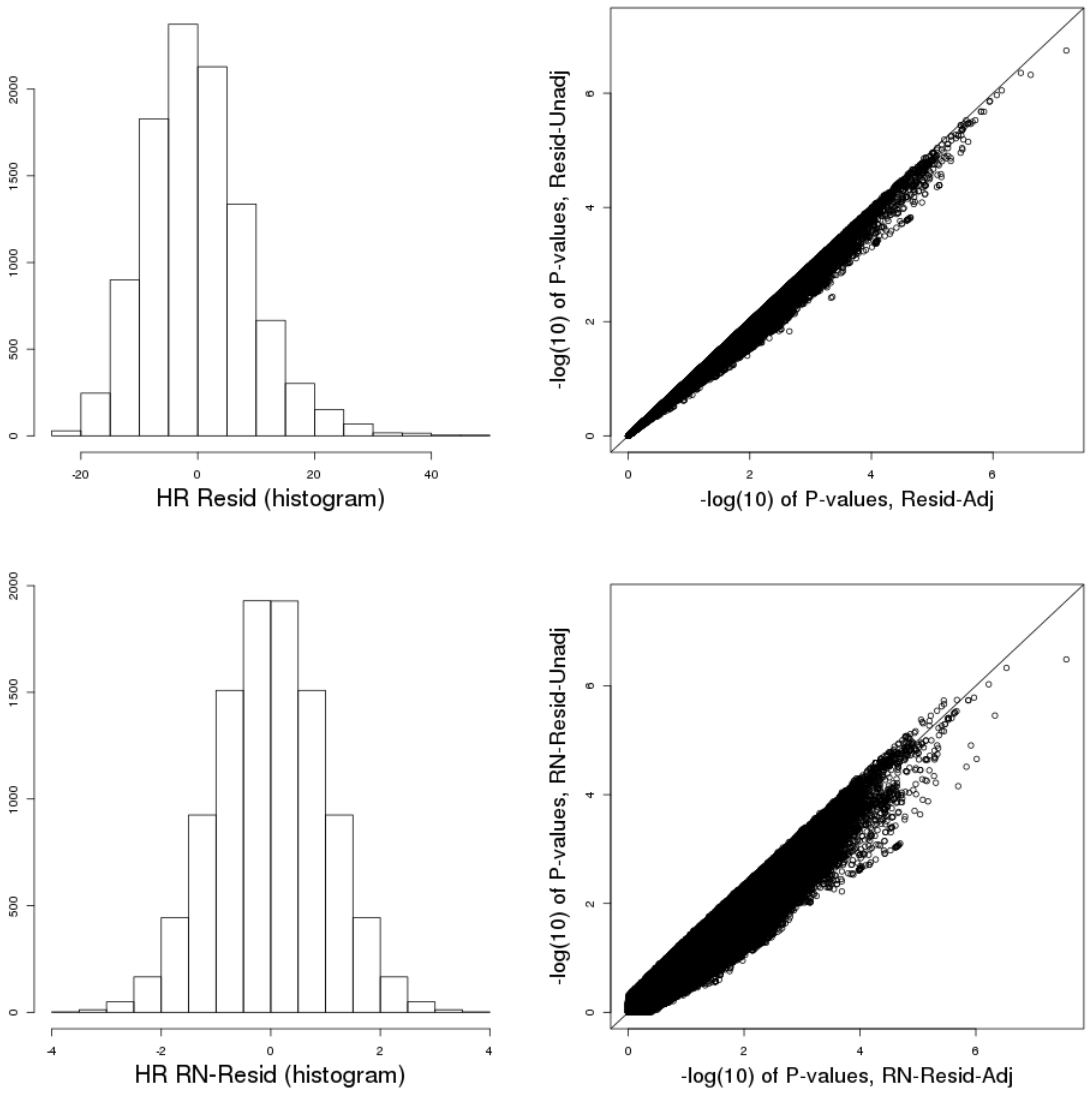


Figure S27: Residuals distribution and GWAS p-values comparisons from the analysis of heart rate in the HCHS/SOL by various approaches.

5.18 QT interval (electrocardiography)

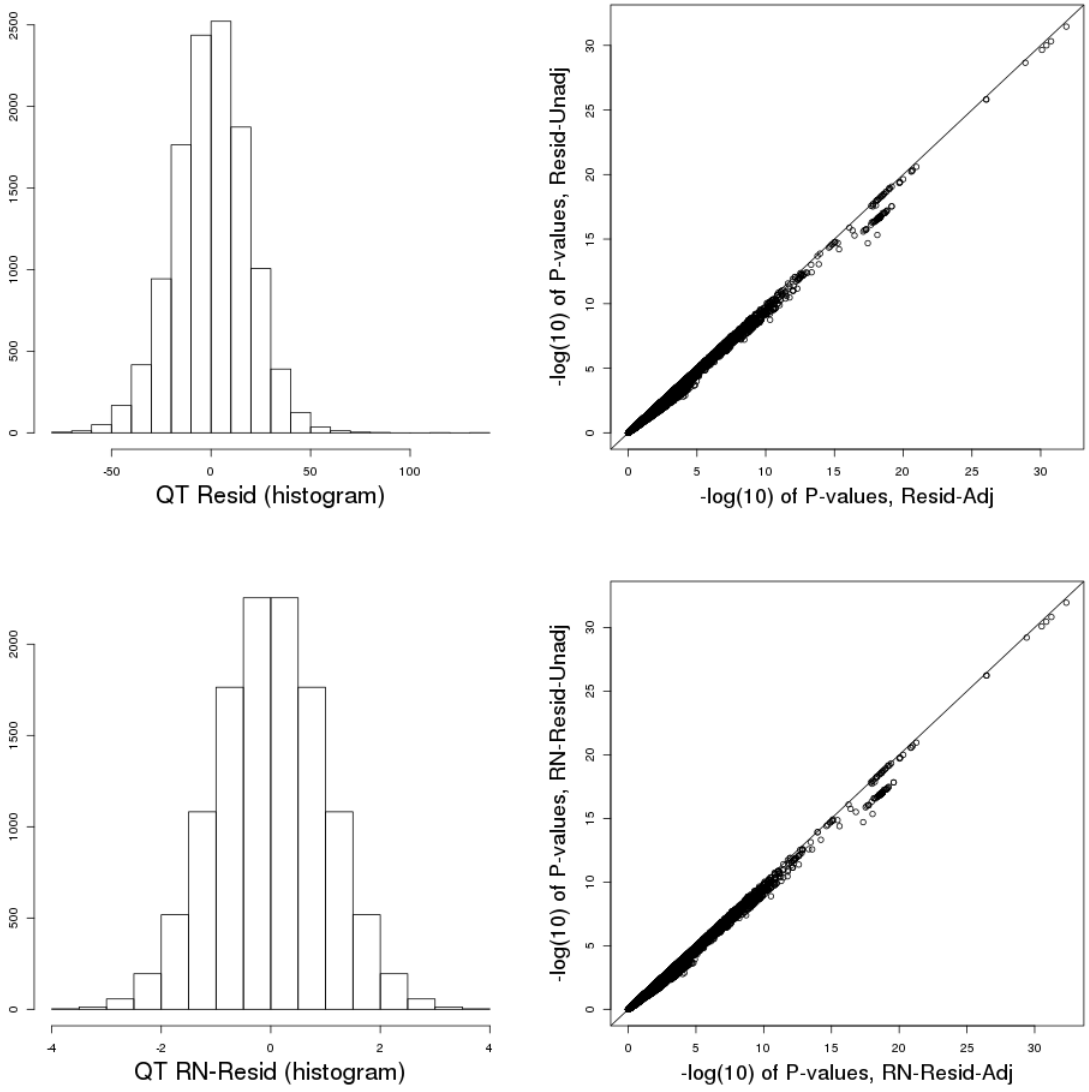


Figure S28: Residuals distribution and GWAS p-values comparisons from the analysis of QT interval length in the HCHS/SOL by various approaches.

5.19 PR interval (electrocardiography)

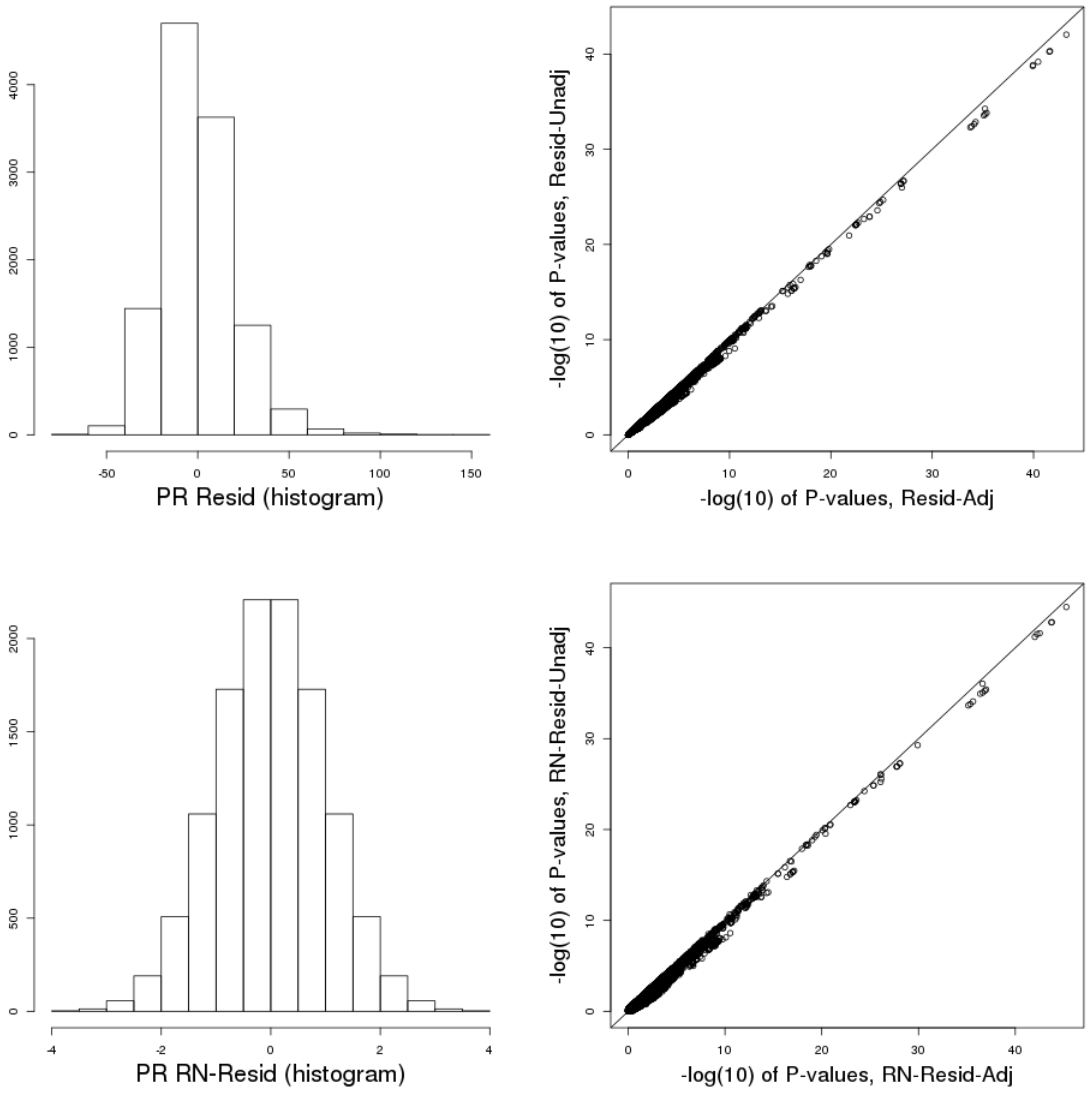


Figure S29: Residuals distribution and GWAS p-values comparisons from the analysis of PR interval length in the HCHS/SOL by various approaches.

5.20 Mean corpuscular hemoglobin concentration

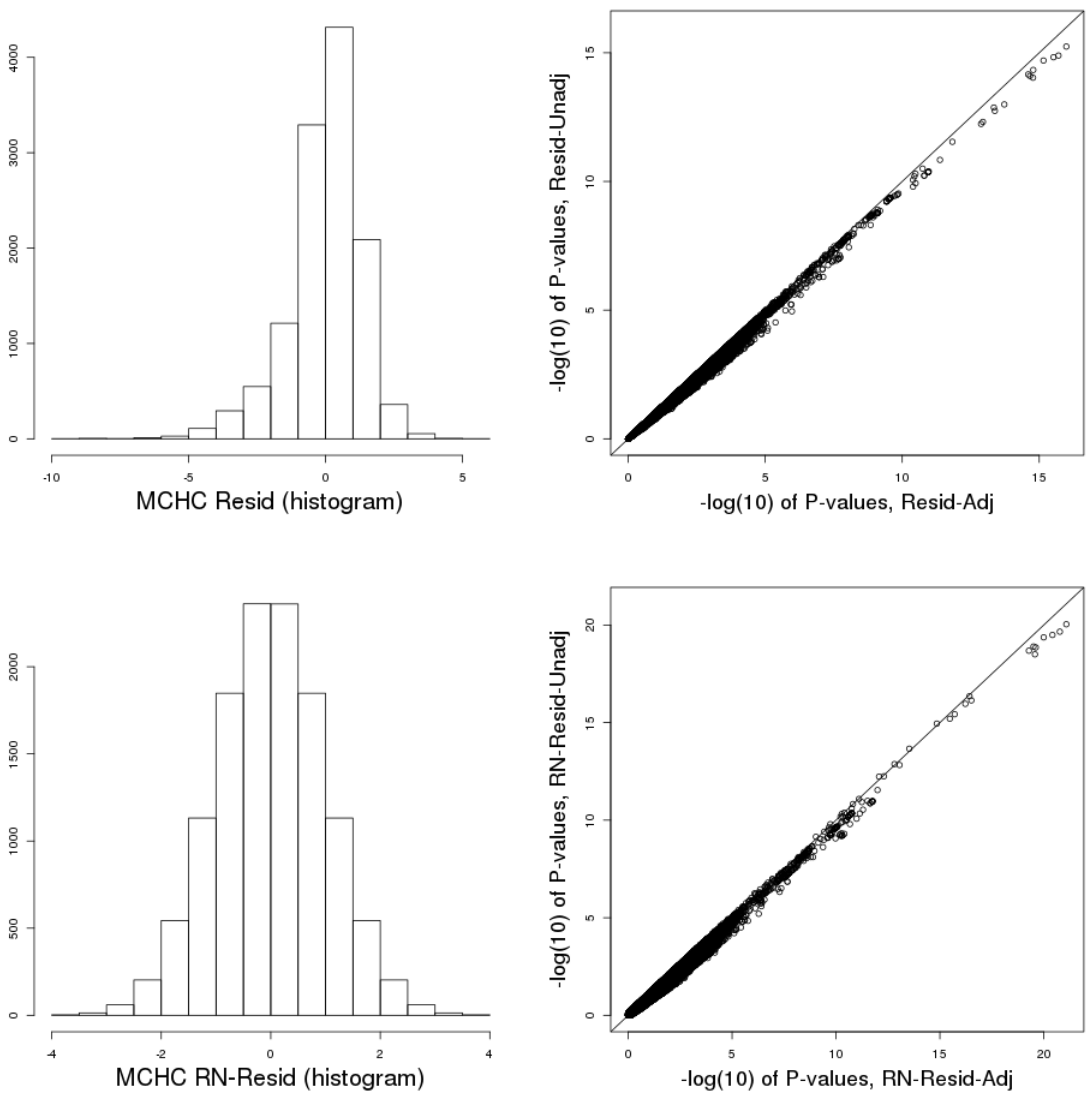


Figure S30: Residuals distribution and GWAS p-values comparisons from the analysis of MCHC in the HCHS/SOL by various approaches.

References

Carpenter, M. A., Crow, R., Steffes, M., Rock, W., Skelton, T., Heilbraun, J., Evans, G., Jensen, R. and Sarpong, D. (2004). Laboratory, reading center, and coordinating center data management methods in the Jackson Heart Study. *The American journal of the medical sciences*, **328** 131–144.

Conomos, M., Laurie, C., Stilp, A., Gogarten, S., McHugh, C., Nelson, S., Sofer, T., Fernandez-Rhodes, L., Justice, A., Graff, M., Young, K., Seyerle, A., Avery, C., Taylor, K., Rotter, J., Talavera, G., Daviglus, M., Wassertheil-Smoller, S., Schneiderman, N., Heiss, G., Kaplan, R., Franceschini, N., Reiner, A., Shaffer, J., Barr, R., Kerr, K., Browning, S., Browning, B., Weir, B., Avilés-Santa, M., Papanicolaou, G., Lumley, T., Szpiro, A., North, K., Rice, K., Thornton, T. and Laurie, C. (2016). Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. *The American Journal of Human Genetics*, **98** 165 – 184.

Delaneau, O., Zagury, J.-F. and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, **10** 5–6.

Dye, B. A., Tan, S., Smith, V., Lewis, B. G., Barker, L. K., Thornton-Evans, G., Eke, P., Beltrán-Aguilar, E., Horowitz, A. and Li, C. (2007). Trends in oral health status: United States, 1988-1994 and 1999-2004. *Vital and health statistics. Series 11, Data from the national health survey* 1–92.

Howie, B. N., Donnelly, P. and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, **5** e1000529.

Laurie, C. et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, **34** 591–602.

LaVange, L. M., Kalsbeek, W. D., Sorlie, P. D., Avilés-Santa, L. M., Kaplan, R. C., Barnhart, J., Liu, K., Giachello, A., Lee, D. J., Ryan, J. et al. (2010). Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Annals of epidemiology*, **20** 642–649.

Sorlie, P. D., Avilés-Santa, L. M., Wassertheil-Smoller, S., Kaplan, R. C., Daviglus, M. L., Giachello, A. L., Schneiderman, N., Raij, L., Talavera, G., Allison, M. et al. (2010). Design and implementation of the hispanic community health study/study of latinos. *Annals of epidemiology*, **20** 629–641.

Wilson, J., Rotimi, C., Ekunwe, L., Royal, C., Crump, M., Wyatt, S., Steffes, M., Adeyemo, A., Zhou, J., Taylor, H. et al. (2005a). The Jackson Heart Study: an overview. *Ethnicity & disease*, **15** S6–S6.

Wilson, J., Rotimi, C., Ekunwe, L., Royal, C., Crump, M., Wyatt, S., Steffes, M., Adeyemo, A., Zhou, J., Taylor, J. H. et al. (2005b). Study design for genetic analysis in the Jackson Heart Study. *Ethnicity & disease*, **15** S6–30.