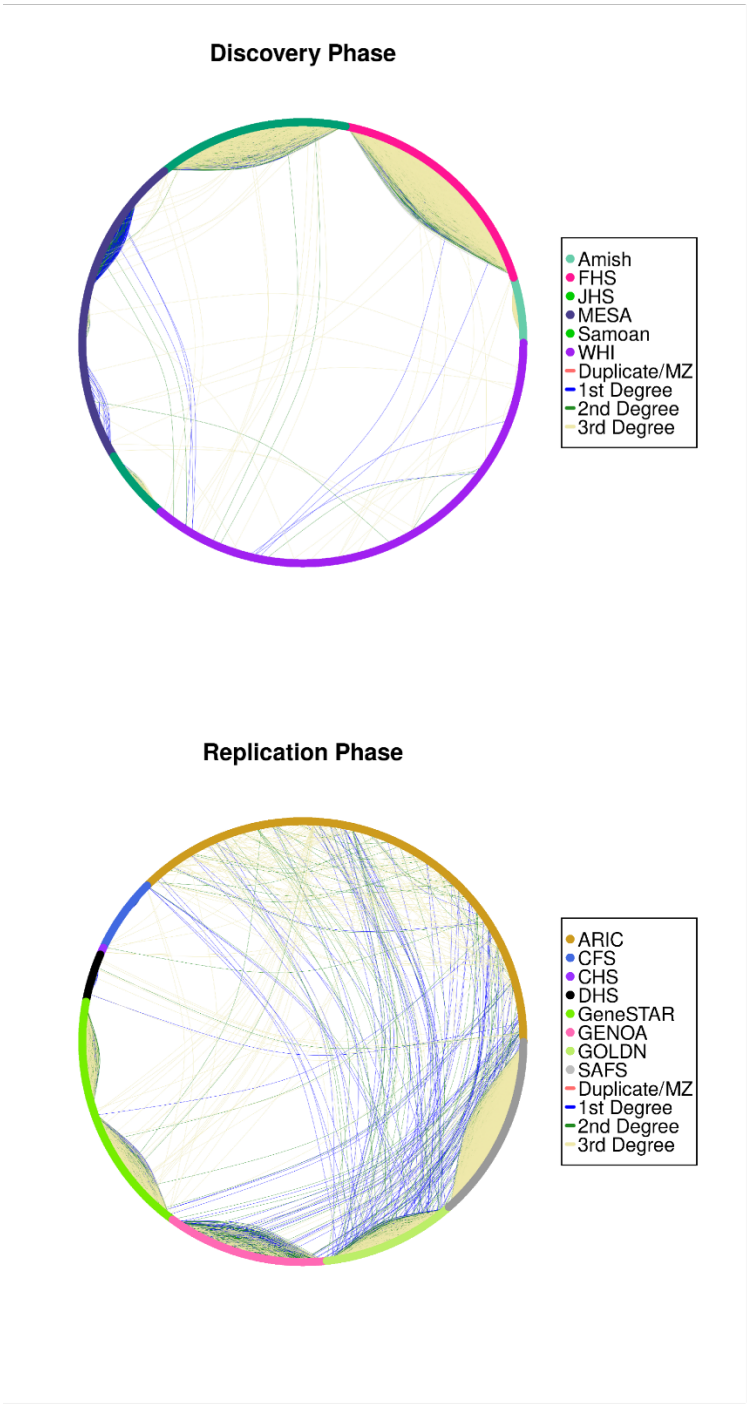
Supplementary Information

A framework for detecting noncoding rare variant associations of large-scale whole-genome sequencing studies

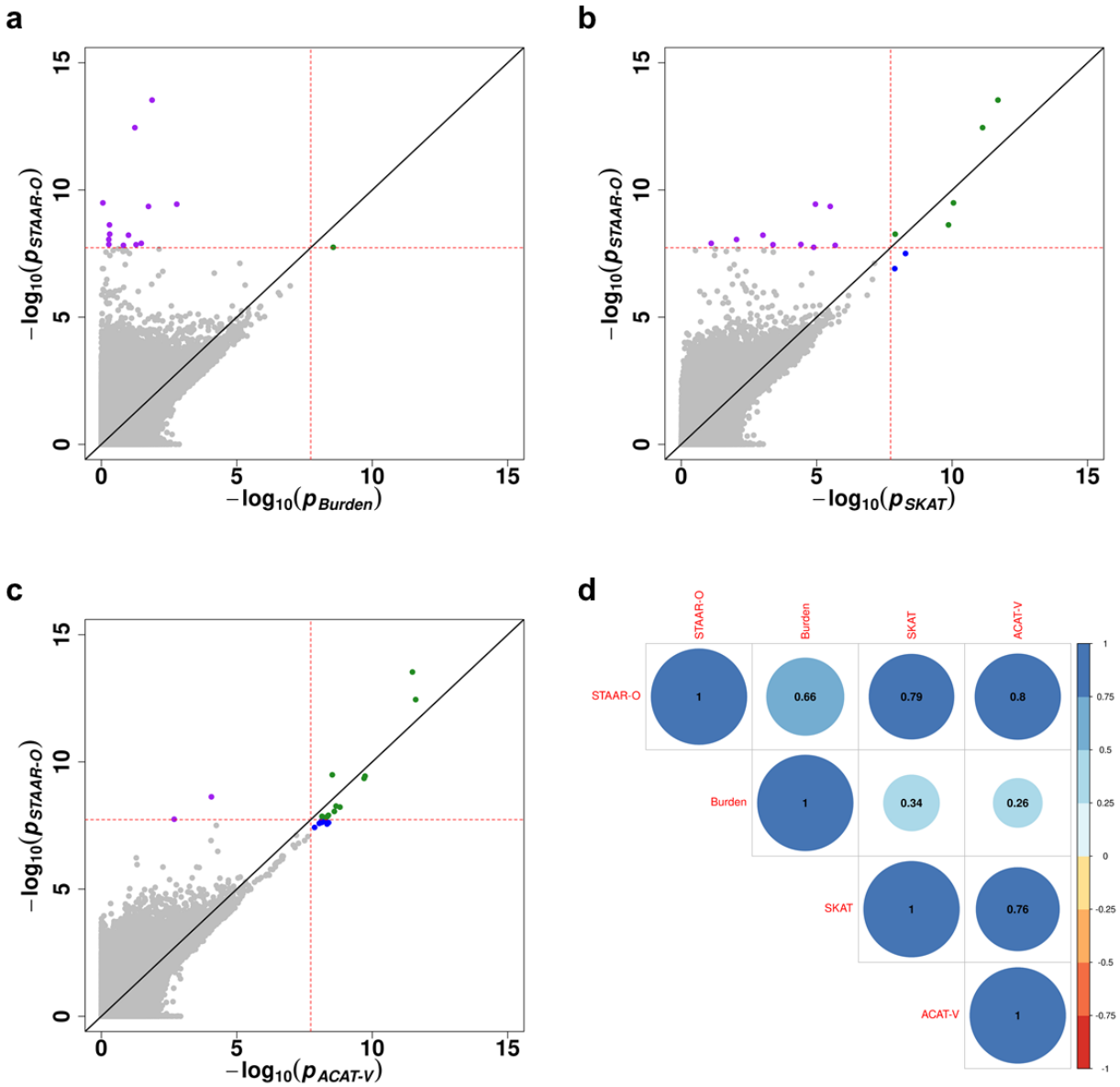
Li et al

Supplementary Figures

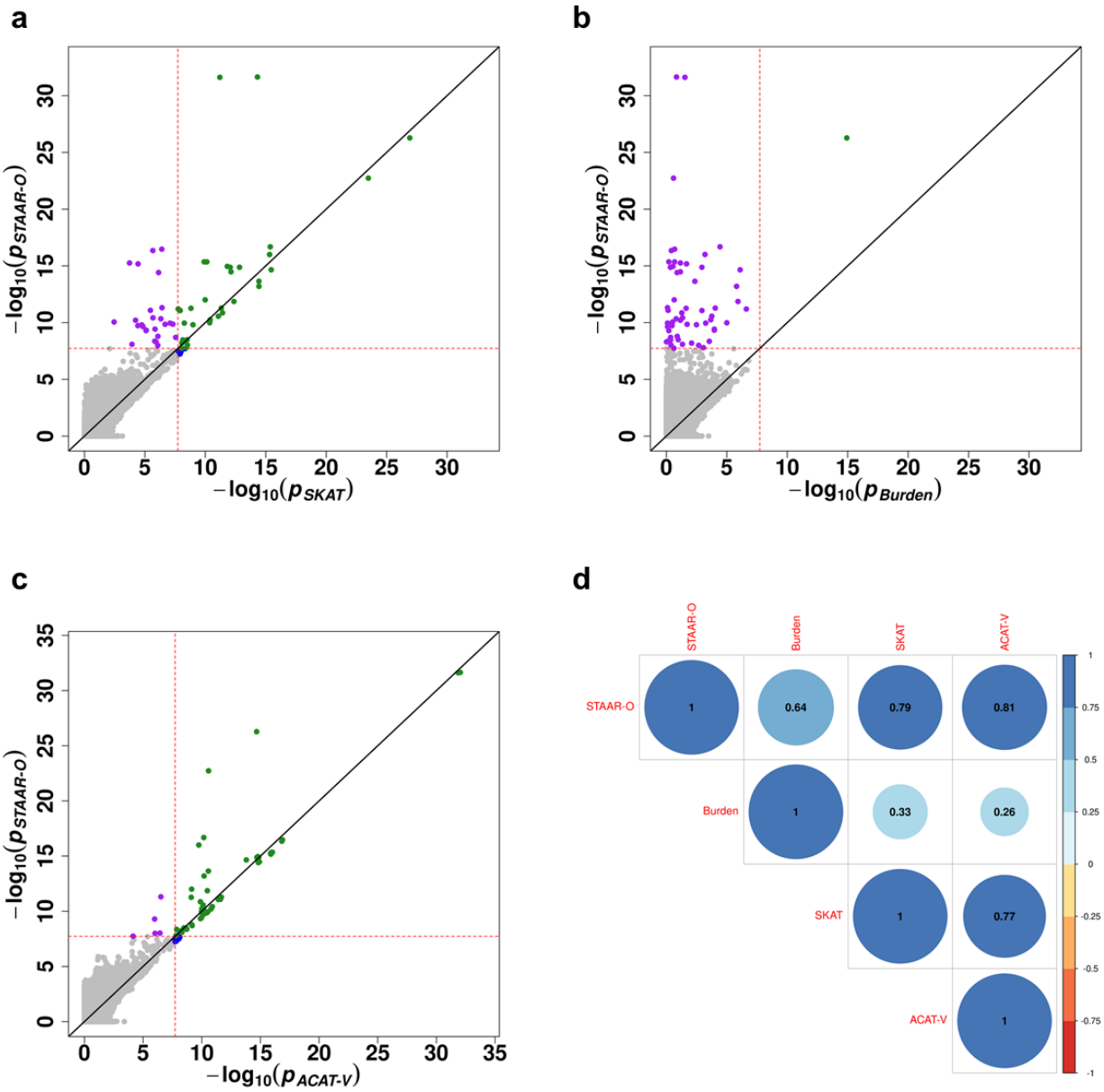
Supplementary Figure 1. Relatedness of subjects within and across studies in the discovery and replication samples of the TOPMed lipids study. See **Supplementary Note** for study abbreviations.



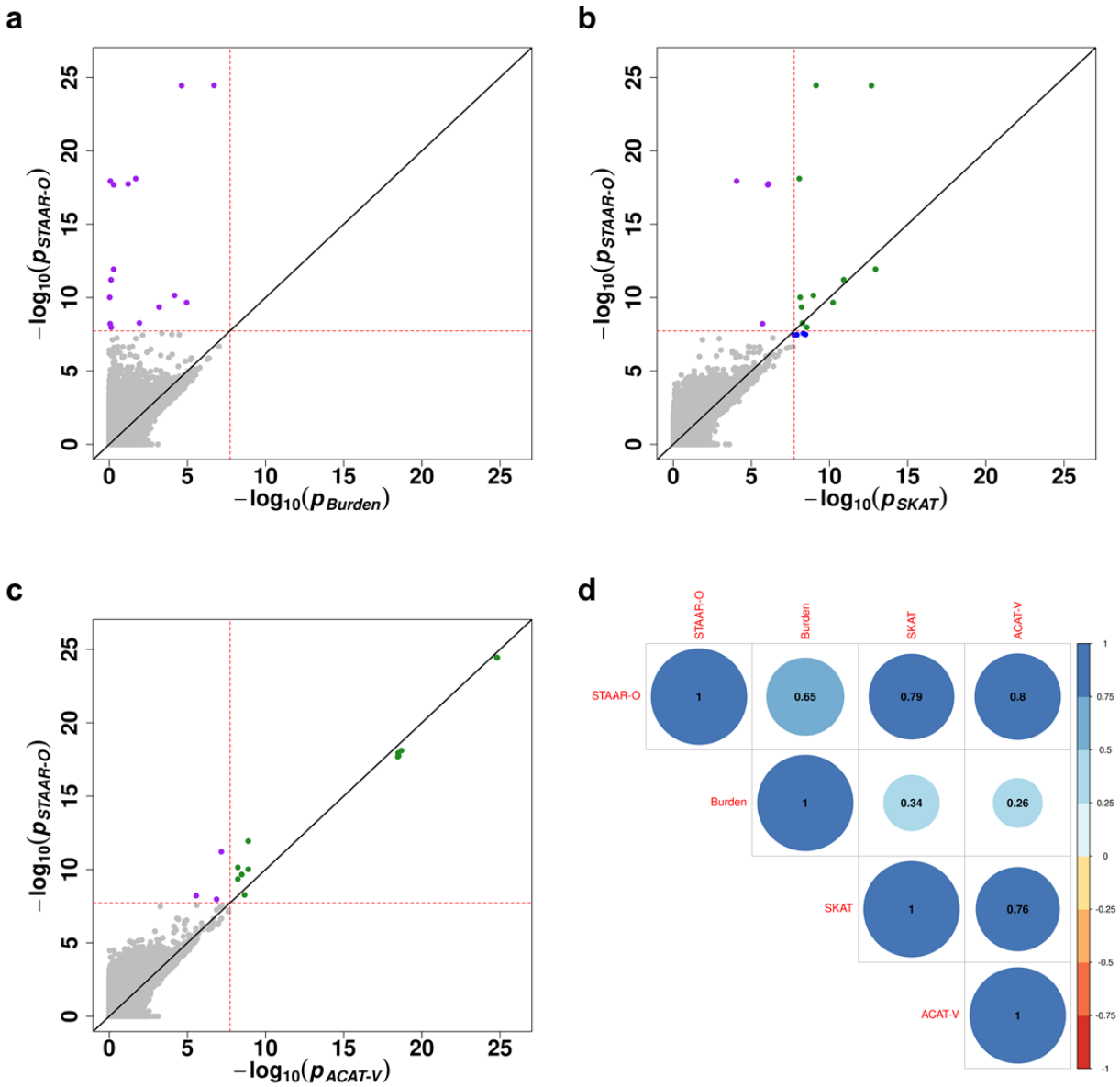
Supplementary Figure 2. Pairwise visual comparisons of the *P* values among the RV tests of the 2-kb sliding window analysis of HDL-C in the discovery phase using the TOPMed cohort (n=21,015). a, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with the burden test. **b**, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with SKAT. **c**, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with the ACAT-V. Each dot represents a sliding window with the x axis label being the $-\log_{10}(P)$ of the conventional test and the y axis label being the $-\log_{10}(P)$ of STAAR-O. **d,** Concordance between the 2-kb sliding window analysis results of different variant-set tests. In panels **a**, **b** and **c**, the colors of the dots indicate the significance of the corresponding sliding windows detected by the variant set tests at the level of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8}$). The green dots indicate the sliding windows are significant under both two tests. The purple dots indicate the sliding window are significant using STAAR-O but are not significant using the other tests (**a**, burden test; **b**, SKAT; **c**, ACAT-V). The blue dots indicate the sliding windows are significant using the other tests (**a**, burden test; **b**, SKAT; **c**, ACAT-V) but are not significant using STAAR-O. The gray dots indicate the sliding window are not significant using both two sets of tests. In all panels, burden test, SKAT, ACAT-V and STAAR-O are two-sided tests.


a
b
c
d
$-\log_{10}(p_{STAAR\text{-}O})$
$-\log_{10}(p_{Burden})$
$-\log_{10}(p_{SKAT})$
$-\log_{10}(p_{ACAT\text{-}V})$
0
5
10
15
STAAR-O
Burden
SKAT
ACAT-V
1
0.66
0.79
0.8
0.34
0.26
0.76
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.75
-1

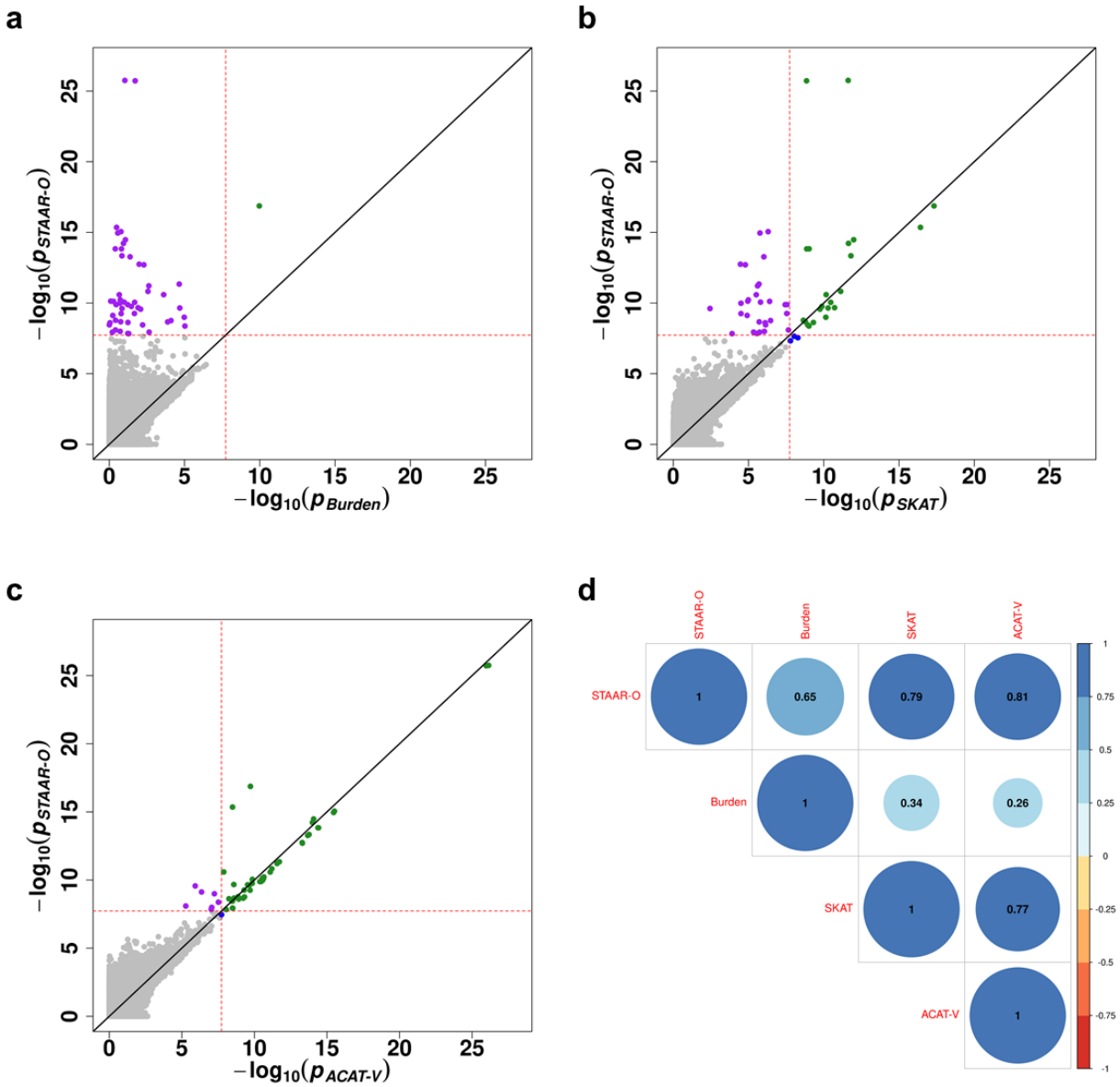
Supplementary Figure 3. Pairwise visual comparisons of the *P* values among the RV tests of the 2-kb sliding window analysis of LDL-C in the discovery phase using the TOPMed cohort (n=21,015). a, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with the burden test. **b**, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with SKAT. **c**, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with the ACAT-V. Each dot represents a sliding window with the x axis label being the $-\log_{10}(P)$ of the conventional test and the y axis label being the $-\log_{10}(P)$ of STAAR-O. **d,** Concordance between the 2-kb sliding window analysis results of different variant-set tests. In panels **a**, **b** and **c**, the colors of the dots indicate the significance of the corresponding sliding windows detected by the variant set tests at the level of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8}$). The green dots indicate the sliding windows are significant under both two tests. The purple dots indicate the sliding window are significant using STAAR-O but are not significant using the other tests (**a**, burden test; **b**, SKAT; **c**, ACAT-V). The blue dots indicate the sliding windows are significant using the other tests (**a**, burden test; **b**, SKAT; **c**, ACAT-V) but are not significant using STAAR-O. The gray dots indicate the sliding window are not significant using both two sets of tests. In all panels, burden test, SKAT, ACAT-V and STAAR-O are two-sided tests.


a
b
c
d
$-\log_{10}(p_{STAAR\text{-}O})$
$-\log_{10}(p_{SKAT})$
$-\log_{10}(p_{Burden})$
$-\log_{10}(p_{ACAT\text{-}V})$
STAAR-O
Burden
SKAT
ACAT-V
1
0.64
0.79
0.81
0.33
0.26
0.77

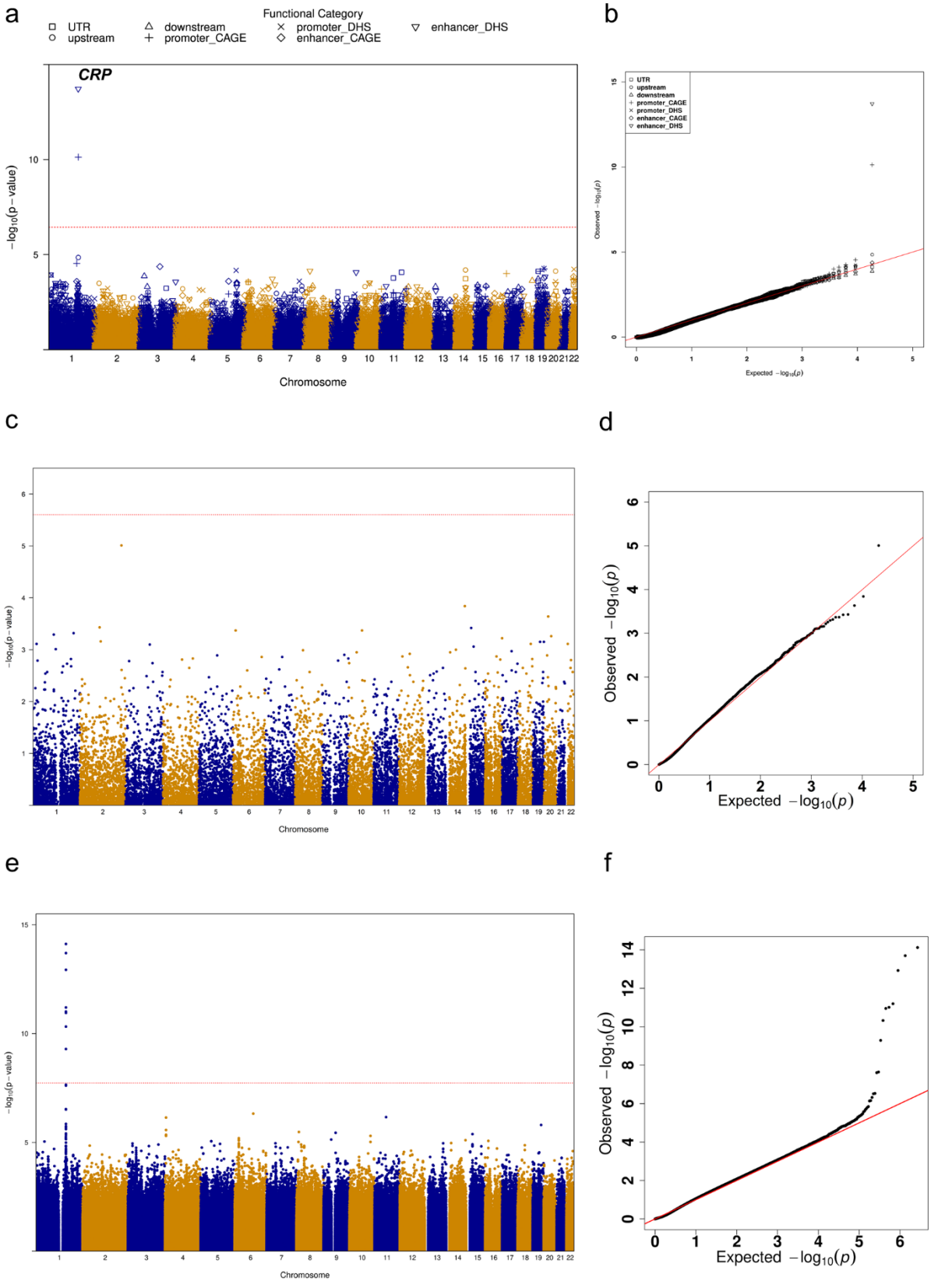
Supplementary Figure 4. Pairwise visual comparisons of the *P* values among the RV tests of the 2-kb sliding window analysis of TG in the discovery phase using the TOPMed cohort (n=21,015). a, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with the burden test. **b**, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with SKAT. **c**, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with the ACAT-V. Each dot represents a sliding window with the x axis label being the $-\log_{10}(P)$ of the conventional test and the y axis label being the $-\log_{10}(P)$ of STAAR-O. **d,** Concordance between the 2-kb sliding window analysis results of different variant-set tests. In panels **a**, **b** and **c**, the colors of the dots indicate the significance of the corresponding sliding windows detected by the variant set tests at the level of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8}$). The green dots indicate the sliding windows are significant under both two tests. The purple dots indicate the sliding window are significant using STAAR-O but are not significant using the other tests (**a**, burden test; **b**, SKAT; **c**, ACAT-V). The blue dots indicate the sliding windows are significant using the other tests (**a**, burden test; **b**, SKAT; **c**, ACAT-V) but are not significant using STAAR-O. The gray dots indicate the sliding window are not significant using both two sets of tests. In all panels, burden test, SKAT, ACAT-V and STAAR-O are two-sided tests.


a
b
c
d
$-\log_{10}(p_{STAAR\text{-}O})$
$-\log_{10}(p_{Burden})$
$-\log_{10}(p_{SKAT})$
$-\log_{10}(p_{ACAT\text{-}V})$
0
5
10
15
20
25
STAAR-O
Burden
SKAT
ACAT-V
1
0.65
0.79
0.8
0.34
0.26
0.76
0.75
0.5
0.25
-0.25
-0.5
-0.75
-1

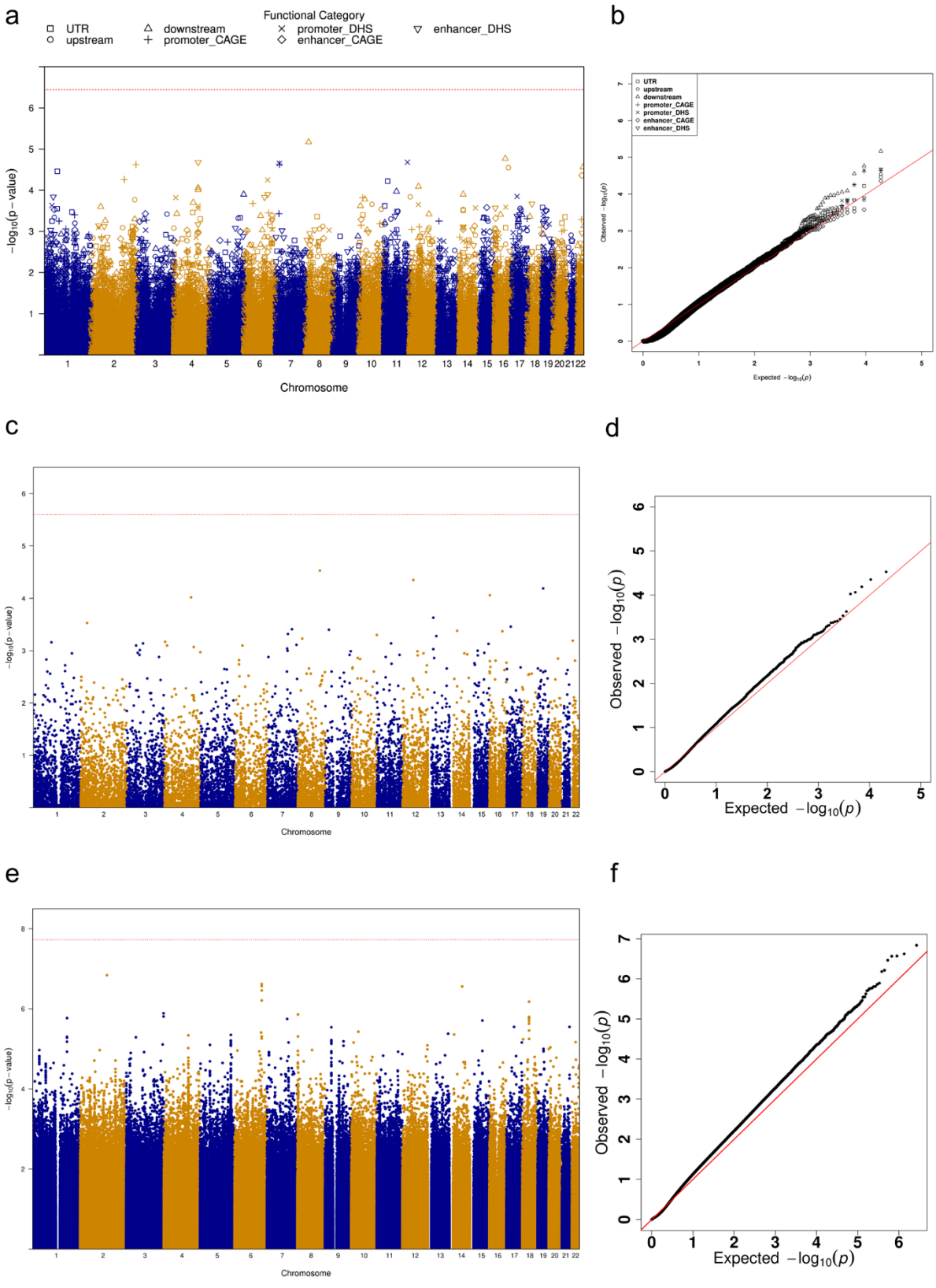
Supplementary Figure 5. Pairwise visual comparisons of the *P* values among the RV tests of the 2-kb sliding window analysis of TC in the discovery phase using the TOPMed cohort (n=21,015). a, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with the burden test. **b**, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with SKAT. **c**, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with the ACAT-V. Each dot represents a sliding window with the x axis label being the $-\log_{10}(P)$ of the conventional test and the y axis label being the $-\log_{10}(P)$ of STAAR-O. **d,** Concordance between the 2-kb sliding window analysis results of different variant-set tests. In panels **a**, **b** and **c**, the colors of the dots indicate the significance of the corresponding sliding windows detected by the variant set tests at the level of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8}$). The green dots indicate the sliding windows are significant under both two tests. The purple dots indicate the sliding window are significant using STAAR-O but are not significant using the other tests (**a**, burden test; **b**, SKAT; **c**, ACAT-V). The blue dots indicate the sliding windows are significant using the other tests (**a**, burden test; **b**, SKAT; **c**, ACAT-V) but are not significant using STAAR-O. The gray dots indicate the sliding window are not significant using both two sets of tests. In all panels, burden test, SKAT, ACAT-V and STAAR-O are two-sided tests.


a
b
c
d
$-\log_{10}(p_{STAAR\text{-}O})$
$-\log_{10}(p_{Burden})$
$-\log_{10}(p_{SKAT})$
$-\log_{10}(p_{ACAT\text{-}V})$
0
5
10
15
20
25
STAAR-O
Burden
SKAT
ACAT-V
1
0.65
0.79
0.81
0.34
0.26
0.77
0.75
0.5
0.25
-0.25
-0.5
-0.75
-1

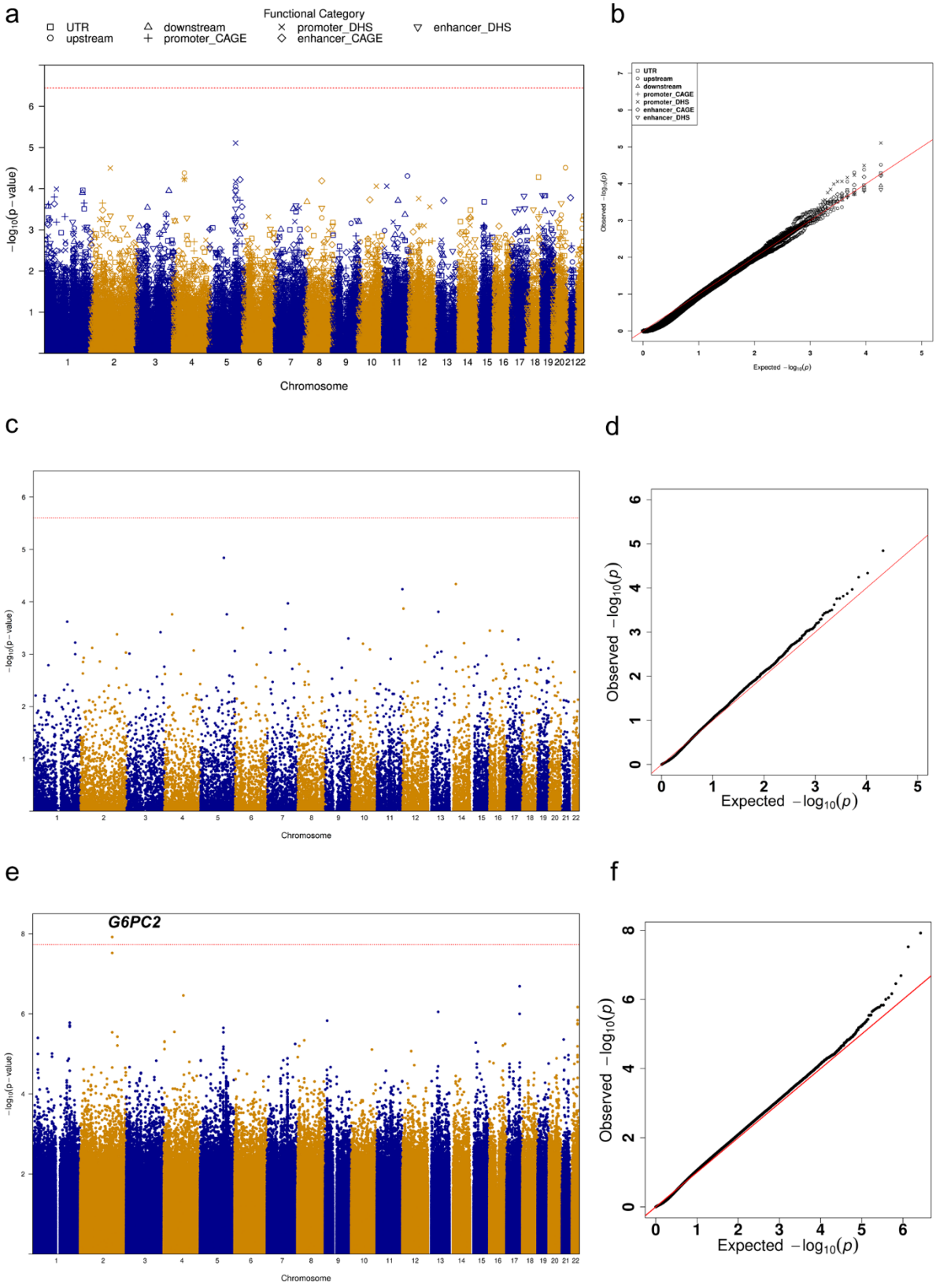
Supplementary Figure 6. Manhattan plots and Q-Q plots for unconditional gene-centric noncoding analysis and sliding window analysis of C-reactive protein (CRP) in the TOPMed Freeze 5 data (n=22,775). a, Manhattan plots for unconditional gene-centric noncoding analysis of protein-coding gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 3.57×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(20{,}000 \times 7) = 3.57 \times 10^{-7}$). Different symbols represent the STAAR-O *P*-value of the protein-coding gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). Promoter_CAGE and promoter_DHS are the promoters with overlap of Cap Analysis of Gene Expression (CAGE) sites and DNase hypersensitivity (DHS) sites for a given gene, respectively. Enhancer_CAGE and enhancer_DHS are the enhancers in GeneHancer predicted regions with the overlap of CAGE sites and DHS sites for a given gene, respectively. **b**, Quantile-quantile plots for unconditional gene-centric noncoding analysis of protein-coding gene. Different symbols represent the STAAR-O *P*-value of the gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). **c,** Manhattan plots for unconditional gene-centric noncoding analysis of ncRNA gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 2.50×10^{-6}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/20{,}000 = 2.50 \times 10^{-6}$). **d,** Quantile-quantile plots for unconditional gene-centric noncoding analysis of ncRNA gene. **e,** Manhattan plot for 2-kb sliding windows. The horizontal line indicates a genome-wide *P*-value threshold of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8}$). **f**, Quantile-quantile plot for 2-kb sliding windows. In panels, **a**, **c** and **e**, the chromosome number are indicated by the colors of dots. In all panels, STAAR-O is a two-sided test.


a
Functional Category
UTR
upstream
downstream
promoter_CAGE
promoter_DHS
enhancer_CAGE
enhancer_DHS
CRP
−log10(p−value)
Chromosome
b
Observed −log10(p)
Expected −log10(p)
c
−log10(p−value)
Chromosome
d
Observed −log10(p)
Expected −log10(p)
e
−log10(p−value)
Chromosome
f
Observed −log10(p)
Expected −log10(p)

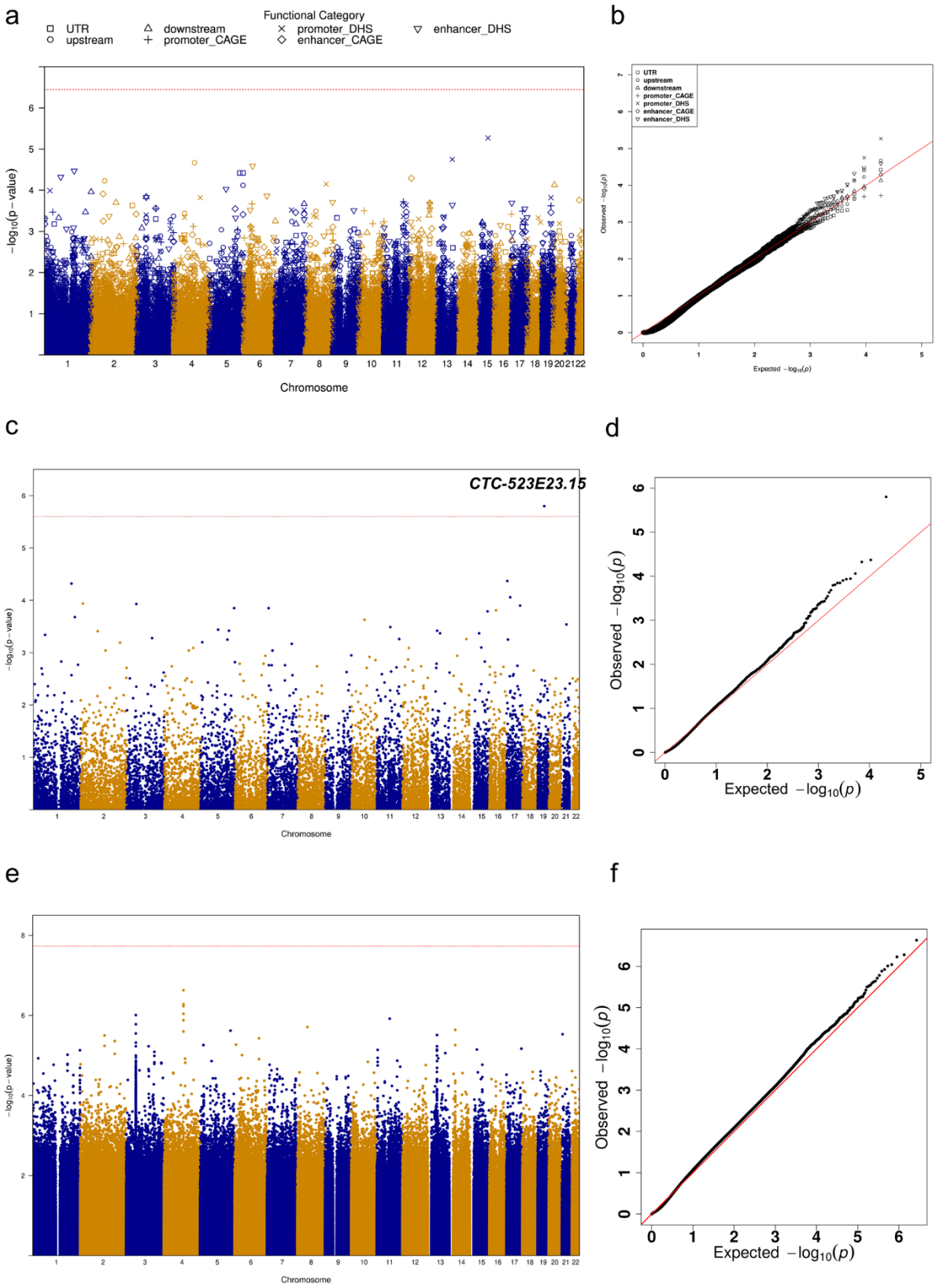
Supplementary Figure 7. Manhattan plots and Q-Q plots for unconditional gene-centric noncoding analysis and sliding window analysis of estimated glomerular filtration rate (eGFR) in the TOPMed Freeze 5 data (n=23,732). a, Manhattan plots for unconditional gene-centric noncoding analysis of protein-coding gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 3.57×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(20{,}000 \times 7) = 3.57 \times 10^{-7}$). Different symbols represent the STAAR-O *P*-value of the protein-coding gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). Promoter_CAGE and promoter_DHS are the promoters with overlap of Cap Analysis of Gene Expression (CAGE) sites and DNase hypersensitivity (DHS) sites for a given gene, respectively. Enhancer_CAGE and enhancer_DHS are the enhancers in GeneHancer predicted regions with the overlap of CAGE sites and DHS sites for a given gene, respectively. **b**, Quantile-quantile plots for unconditional gene-centric noncoding analysis of protein-coding gene. Different symbols represent the STAAR-O *P*-value of the gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). **c,** Manhattan plots for unconditional gene-centric noncoding analysis of ncRNA gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 2.50×10^{-6}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/20{,}000 = 2.50 \times 10^{-6}$). **d,** Quantile-quantile plots for unconditional gene-centric noncoding analysis of ncRNA gene. **e,** Manhattan plot for 2-kb sliding windows. The horizontal line indicates a genome-wide *P*-value threshold of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8}$). **f**, Quantile-quantile plot for 2-kb sliding windows. In panels, **a**, **c** and **e**, the chromosome number are indicated by the colors of dots. In all panels, STAAR-O is a two-sided test.


a
Functional Category
UTR
upstream
downstream
promoter_CAGE
promoter_DHS
enhancer_CAGE
enhancer_DHS
−log10(p−value)
Chromosome
b
Observed −log10(p)
Expected −log10(p)
c
−log10(p−value)
Chromosome
d
Observed −log10(p)
Expected −log10(p)
e
−log10(p−value)
Chromosome
f
Observed −log10(p)
Expected −log10(p)

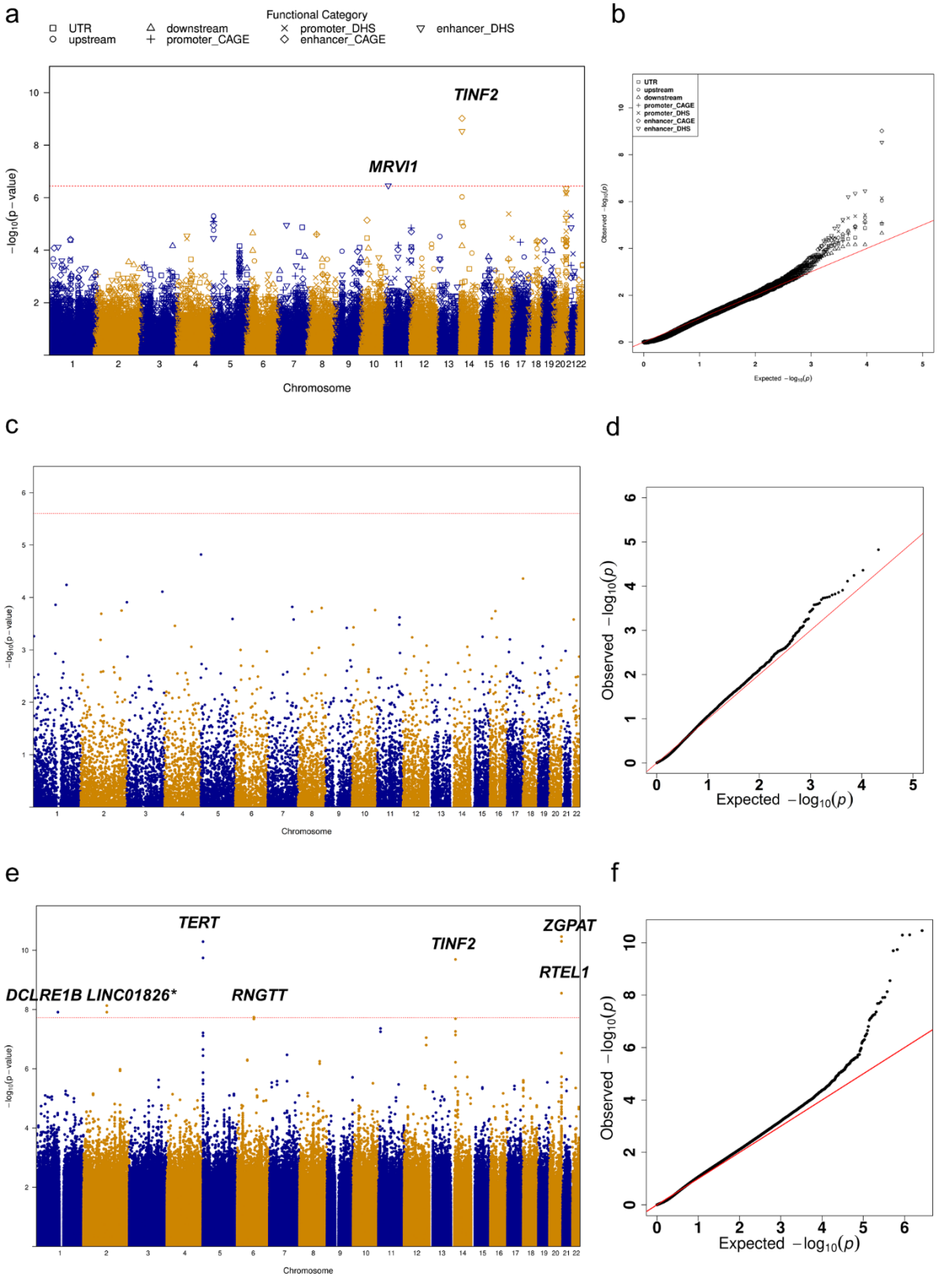
Supplementary Figure 8. Manhattan plots and Q-Q plots for unconditional gene-centric noncoding analysis and sliding window analysis of fasting glucose level (FG) in the TOPMed Freeze 5 data (n=23,859). a, Manhattan plots for unconditional gene-centric noncoding analysis of protein-coding gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 3.57×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(20{,}000 \times 7) = 3.57 \times 10^{-7}$). Different symbols represent the STAAR-O *P*-value of the protein-coding gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). Promoter_CAGE and promoter_DHS are the promoters with overlap of Cap Analysis of Gene Expression (CAGE) sites and DNase hypersensitivity (DHS) sites for a given gene, respectively. Enhancer_CAGE and enhancer_DHS are the enhancers in GeneHancer predicted regions with the overlap of CAGE sites and DHS sites for a given gene, respectively. **b**, Quantile-quantile plots for unconditional gene-centric noncoding analysis of protein-coding gene. Different symbols represent the STAAR-O *P*-value of the gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). **c,** Manhattan plots for unconditional gene-centric noncoding analysis of ncRNA gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 2.50×10^{-6}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/20{,}000 = 2.50 \times 10^{-6}$). **d,** Quantile-quantile plots for unconditional gene-centric noncoding analysis of ncRNA gene. **e,** Manhattan plot for 2-kb sliding windows. The horizontal line indicates a genome-wide *P*-value threshold of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8}$). **f**, Quantile-quantile plot for 2-kb sliding windows. In panels, **a**, **c** and **e**, the chromosome number are indicated by the colors of dots. In all panels, STAAR-O is a two-sided test.


a
Functional Category
UTR
upstream
downstream
promoter_CAGE
promoter_DHS
enhancer_CAGE
enhancer_DHS
−log10(p−value)
Chromosome
b
Observed −log10(p)
Expected −log10(p)
c
−log10(p−value)
Chromosome
d
Observed −log10(p)
Expected −log10(p)
e
G6PC2
−log10(p−value)
Chromosome
f
Observed −log10(p)
Expected −log10(p)

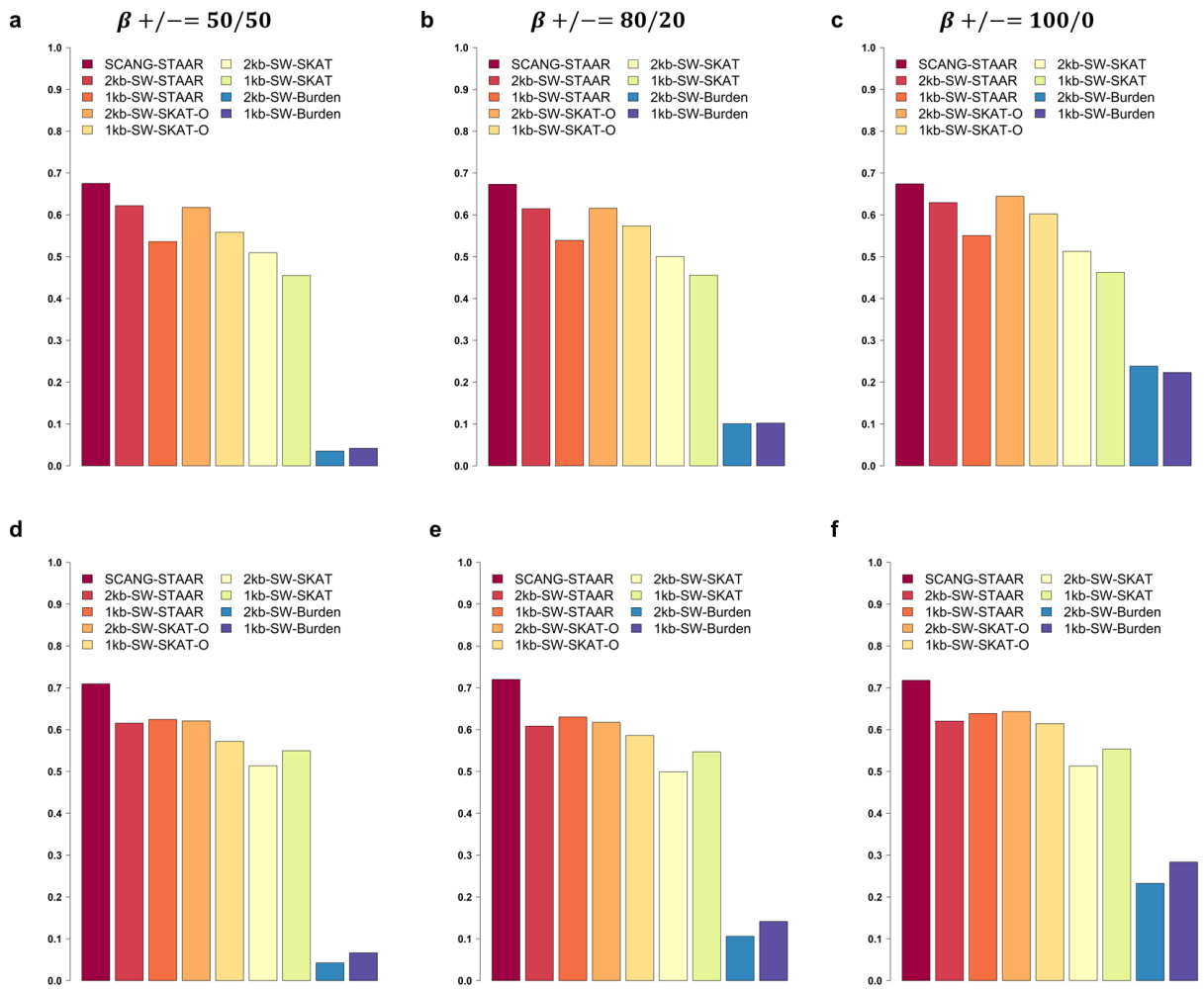
Supplementary Figure 9. Manhattan plots and Q-Q plots for unconditional gene-centric noncoding analysis and sliding window analysis of fasting insulin level (FI) in the TOPMed Freeze 5 data (n=21,900). a, Manhattan plots for unconditional gene-centric noncoding analysis of protein-coding gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 3.57×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(20{,}000 \times 7) = 3.57 \times 10^{-7}$). Different symbols represent the STAAR-O *P*-value of the protein-coding gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). Promoter_CAGE and promoter_DHS are the promoters with overlap of Cap Analysis of Gene Expression (CAGE) sites and DNase hypersensitivity (DHS) sites for a given gene, respectively. Enhancer_CAGE and enhancer_DHS are the enhancers in GeneHancer predicted regions with the overlap of CAGE sites and DHS sites for a given gene, respectively. **b**, Quantile-quantile plots for unconditional gene-centric noncoding analysis of protein-coding gene. Different symbols represent the STAAR-O *P*-value of the gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). **c,** Manhattan plots for unconditional gene-centric noncoding analysis of ncRNA gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 2.50×10^{-6}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/20{,}000 = 2.50 \times 10^{-6}$). **d,** Quantile-quantile plots for unconditional gene-centric noncoding analysis of ncRNA gene. **e,** Manhattan plot for 2-kb sliding windows. The horizontal line indicates a genome-wide *P*-value threshold of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8}$). **f**, Quantile-quantile plot for 2-kb sliding windows. In panels, **a**, **c** and **e**, the chromosome number are indicated by the colors of dots. In all panels, STAAR-O is a two-sided test.


a
Functional Category
UTR
upstream
downstream
promoter_CAGE
promoter_DHS
enhancer_CAGE
enhancer_DHS
−log10(p−value)
Chromosome
b
Observed −log10(p)
Expected −log10(p)
c
CTC-523E23.15
d
e
f

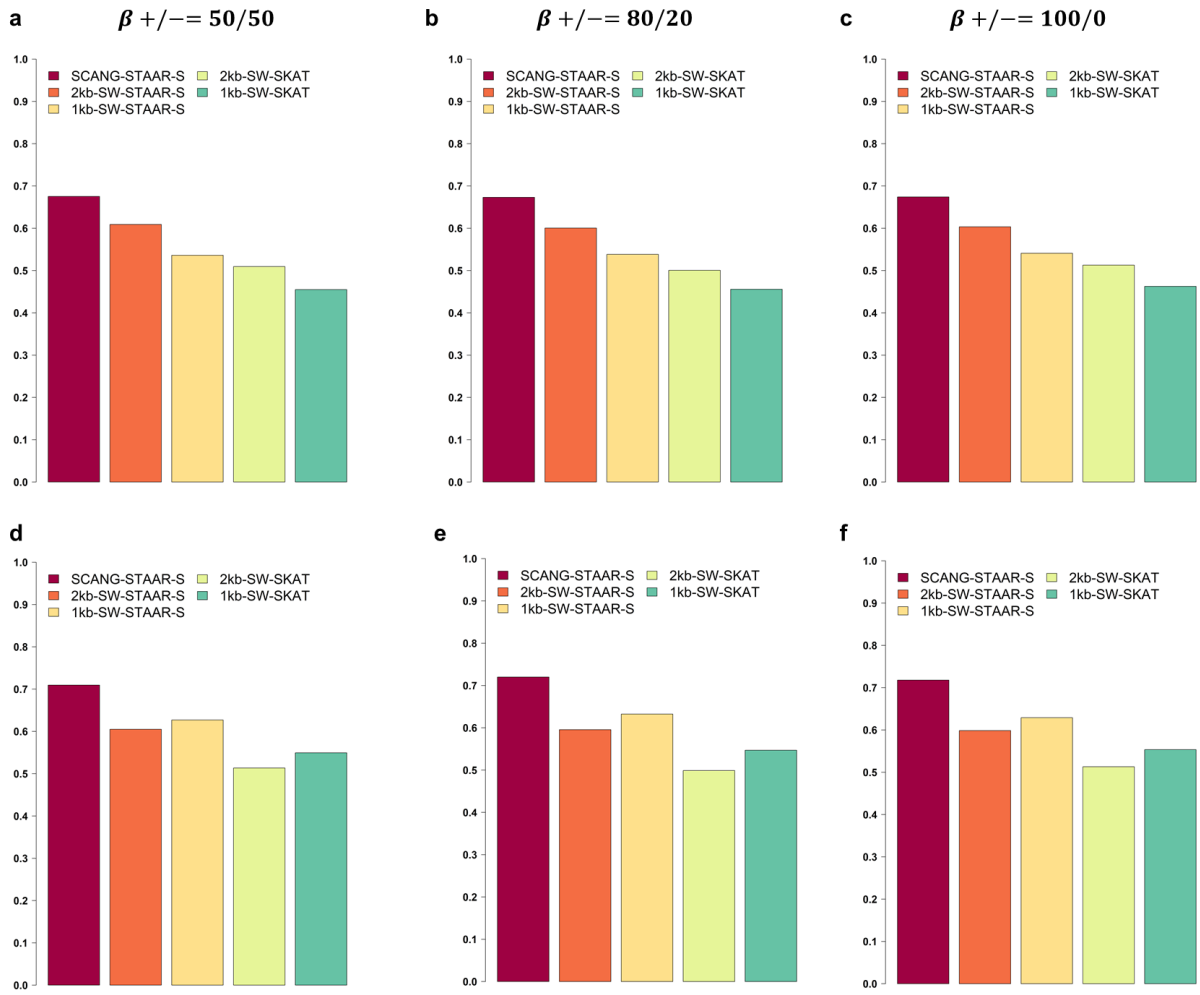
Supplementary Figure 10. Manhattan plots and Q-Q plots for unconditional gene-centric noncoding analysis and sliding window analysis of telomere length (TL) in the TOPMed Freeze 5 data (n=39,742). a, Manhattan plots for unconditional gene-centric noncoding analysis of protein-coding gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 3.57×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(20{,}000 \times 7) = 3.57 \times 10^{-7}$). Different symbols represent the STAAR-O *P*-value of the protein-coding gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). Promoter_CAGE and promoter_DHS are the promoters with overlap of Cap Analysis of Gene Expression (CAGE) sites and DNase hypersensitivity (DHS) sites for a given gene, respectively. Enhancer_CAGE and enhancer_DHS are the enhancers in GeneHancer predicted regions with the overlap of CAGE sites and DHS sites for a given gene, respectively. **b**, Quantile-quantile plots for unconditional gene-centric noncoding analysis of protein-coding gene. Different symbols represent the STAAR-O *P*-value of the gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). **c,** Manhattan plots for unconditional gene-centric noncoding analysis of ncRNA gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 2.50×10^{-6}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/20{,}000 = 2.50 \times 10^{-6}$). **d,** Quantile-quantile plots for unconditional gene-centric noncoding analysis of ncRNA gene. **e,** Manhattan plot for 2-kb sliding windows. The horizontal line indicates a genome-wide *P*-value threshold of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8}$). **f**, Quantile-quantile plot for 2-kb sliding windows. In panels, **a**, **c** and **e**, the chromosome number are indicated by the colors of dots. In all panels, STAAR-O is a two-sided test.


a
Functional Category
UTR
upstream
downstream
promoter_CAGE
promoter_DHS
enhancer_CAGE
enhancer_DHS
TINF2
MRVI1
−log10(p−value)
Chromosome
b
UTR
upstream
downstream
promoter_CAGE
promoter_DHS
enhancer_CAGE
enhancer_DHS
Observed −log10(p)
Expected −log10(p)
c
−log10(p−value)
Chromosome
d
Observed −log10(p)
Expected −log10(p)
e
TERT
TINF2
ZGPAT
RTEL1
DCLRE1B LINC01826*
RNGTT
−log10(p−value)
Chromosome
f
Observed −log10(p)
Expected −log10(p)

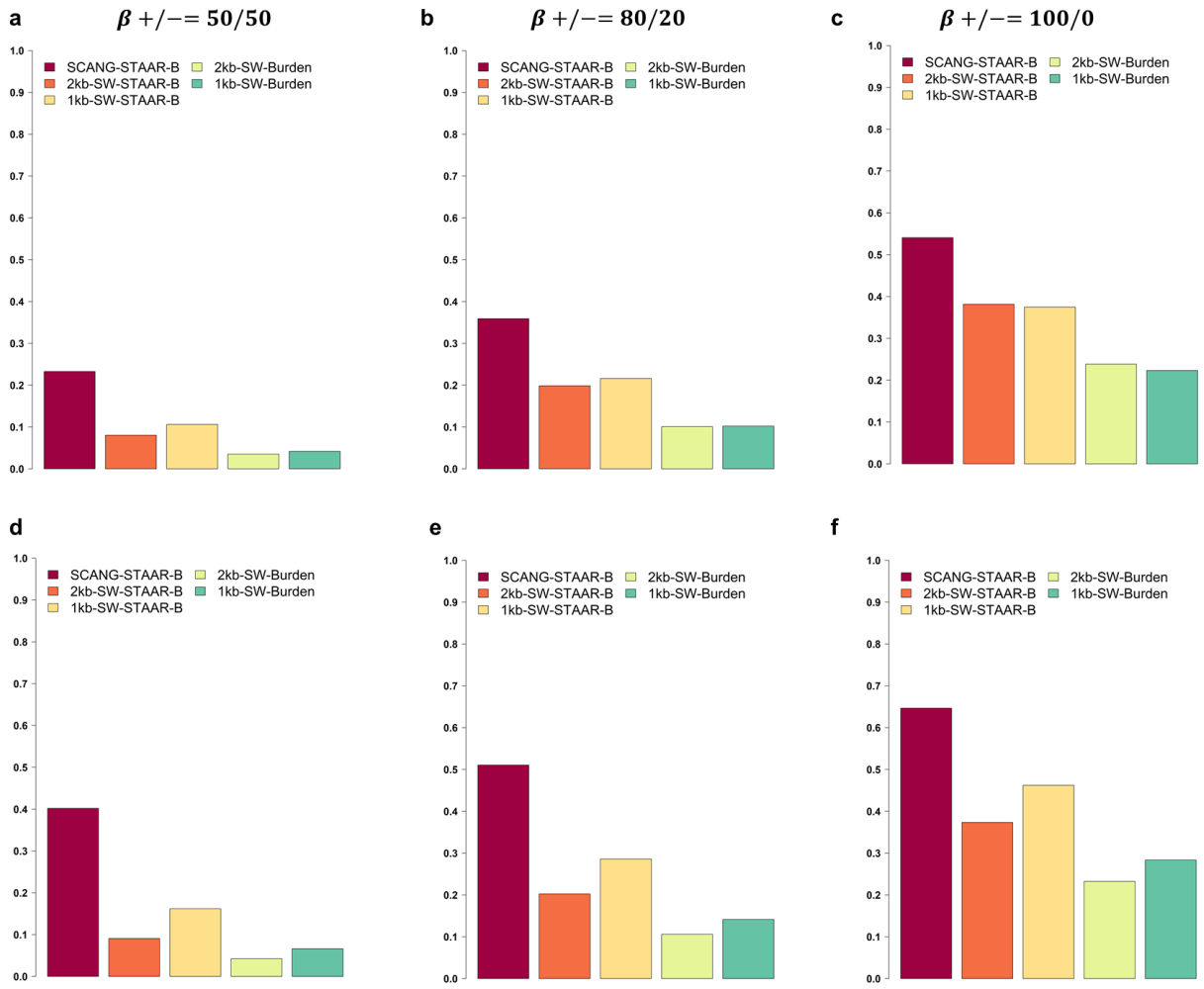
Supplementary Figure 11. Power comparisons of dynamic window procedure SCANG-STAAR and sliding window procedures using burden, SKAT, SKAT-O, and STAAR for continuous trait analysis. Empirical power was evaluated by the causal variants detection rate and the signal region detection rate defined in the simulation section. Both criteria were calculated at the genome-wise/family-wise type I error $\alpha = 0.01$. We randomly selected two signal regions (variant phenotype association regions) across the 10-Mb genome in each simulation replicate. The length of the signal regions was randomly selected from 1 kb, 1.5 kb, and 2 kb. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model, and on average there were 15% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0|\log_{10} MAF_j|$, where c_0 was set to be 0.10. From left to right, the plots consider settings in which the effect sizes for the causal variants are 50% positive (50% negative), 80% positive (20% negative) and 100% positive (0% negative). The causal variants detection rate (A-C) is the proportion of detected causal variants in 1,000 simulated whole-genome data sets. A causal variant is defined as detected if it is in one of the detected signal regions. The signal region detection rate (D-F) is the proportion of detected signal regions in 1,000 simulated whole-genome data sets. A signal region is defined as detected if it overlapped with one of the detected signal regions. For each configuration, the sample size was 50,000. For each setting, nine methods were compared: SCANG-STAAR-S and 1-kb and 2-kb sliding window procedures (SW for short) using burden, SKAT, SKAT-O and STAAR-O. For SCANG-STAAR-S, the range of search window lengths was set by the number of variants in searching windows between 40 and 300.


a
β +/−= 50/50
b
β +/−= 80/20
c
β +/−= 100/0
d
e
f
SCANG-STAAR
2kb-SW-STAAR
1kb-SW-STAAR
2kb-SW-SKAT-O
1kb-SW-SKAT-O
2kb-SW-SKAT
1kb-SW-SKAT
2kb-SW-Burden
1kb-SW-Burden
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

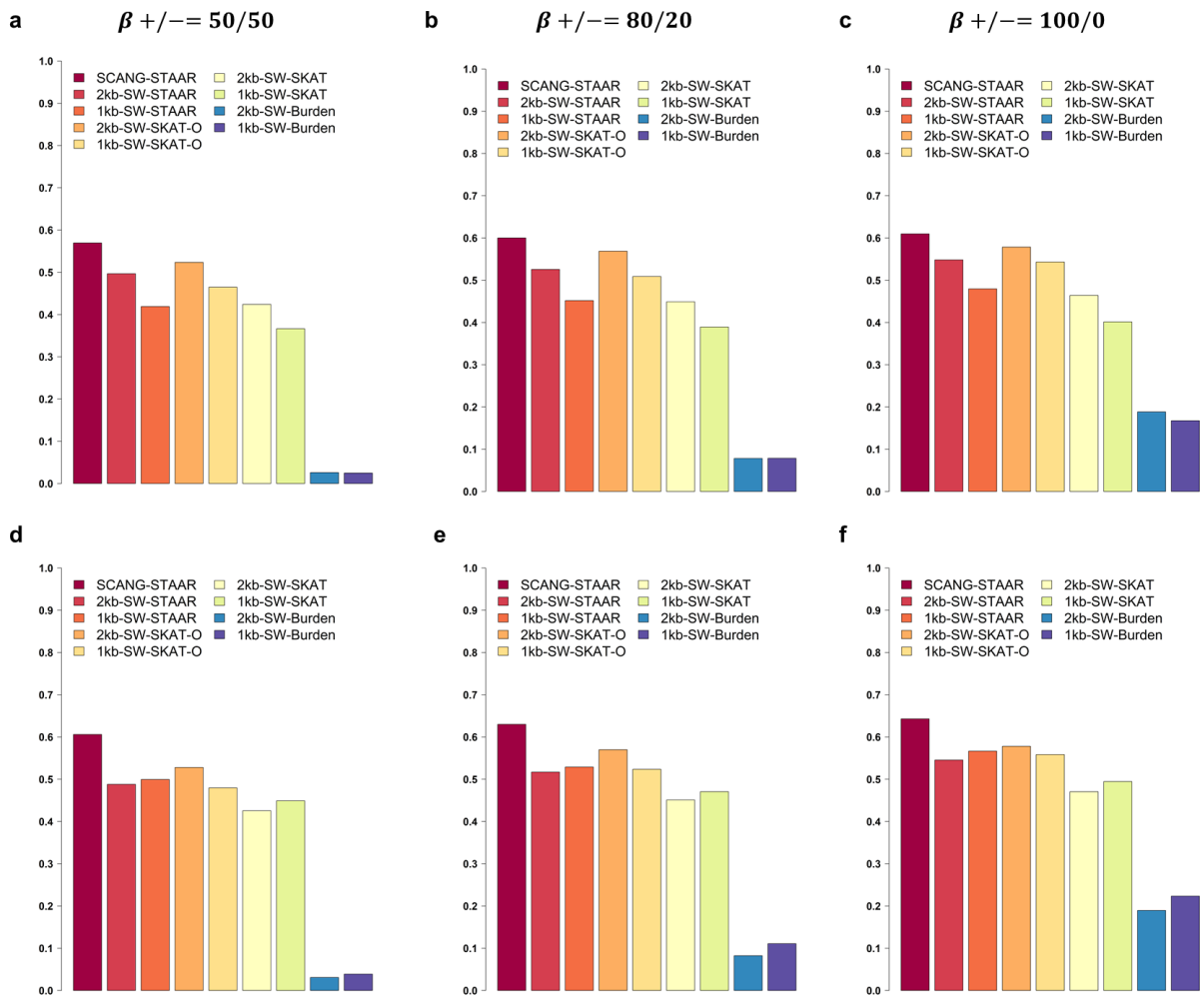
Supplementary Figure 12. Power comparisons of dynamic window procedure SCANG-STAAR-S and the corresponding sliding window procedures using SKAT for continuous trait analysis. Empirical power was evaluated by the causal variants detection rate and the signal region detection rate defined in the simulation section. Both criteria were calculated at the genome-wise/family-wise type I error $\alpha = 0.01$. We randomly selected two signal regions (variant phenotype association regions) across the 10-Mb genome in each simulation replicate. The length of the signal regions was randomly selected from 1 kb, 1.5 kb, and 2 kb. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model, and on average there were 15% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$, where c_0 was set to be 0.10. From left to right, the plots consider settings in which the effect sizes for the causal variants are 50% positive (50% negative), 80% positive (20% negative) and 100% positive (0% negative). The causal variants detection rate (A-C) is the proportion of detected causal variants in 1,000 simulated whole-genome data sets. A causal variant is defined as detected if it is in one of the detected signal regions. The signal region detection rate (D-F) is the proportion of detected signal regions in 1,000 simulated whole-genome data sets. A signal region is defined as detected if it overlapped with one of the detected signal regions. For each configuration, the sample size was 50,000. For each setting, five methods were compared: SCANG-STAAR-S and 1-kb and 2-kb sliding window procedures (SW for short) using SKAT and STAAR-S. For SCANG-STAAR-S, the range of search window lengths was set by the number of variants in searching windows between 40 and 300.


a
β +/−= 50/50
b
β +/−= 80/20
c
β +/−= 100/0
d
e
f
SCANG-STAAR-S
2kb-SW-STAAR-S
1kb-SW-STAAR-S
2kb-SW-SKAT
1kb-SW-SKAT
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

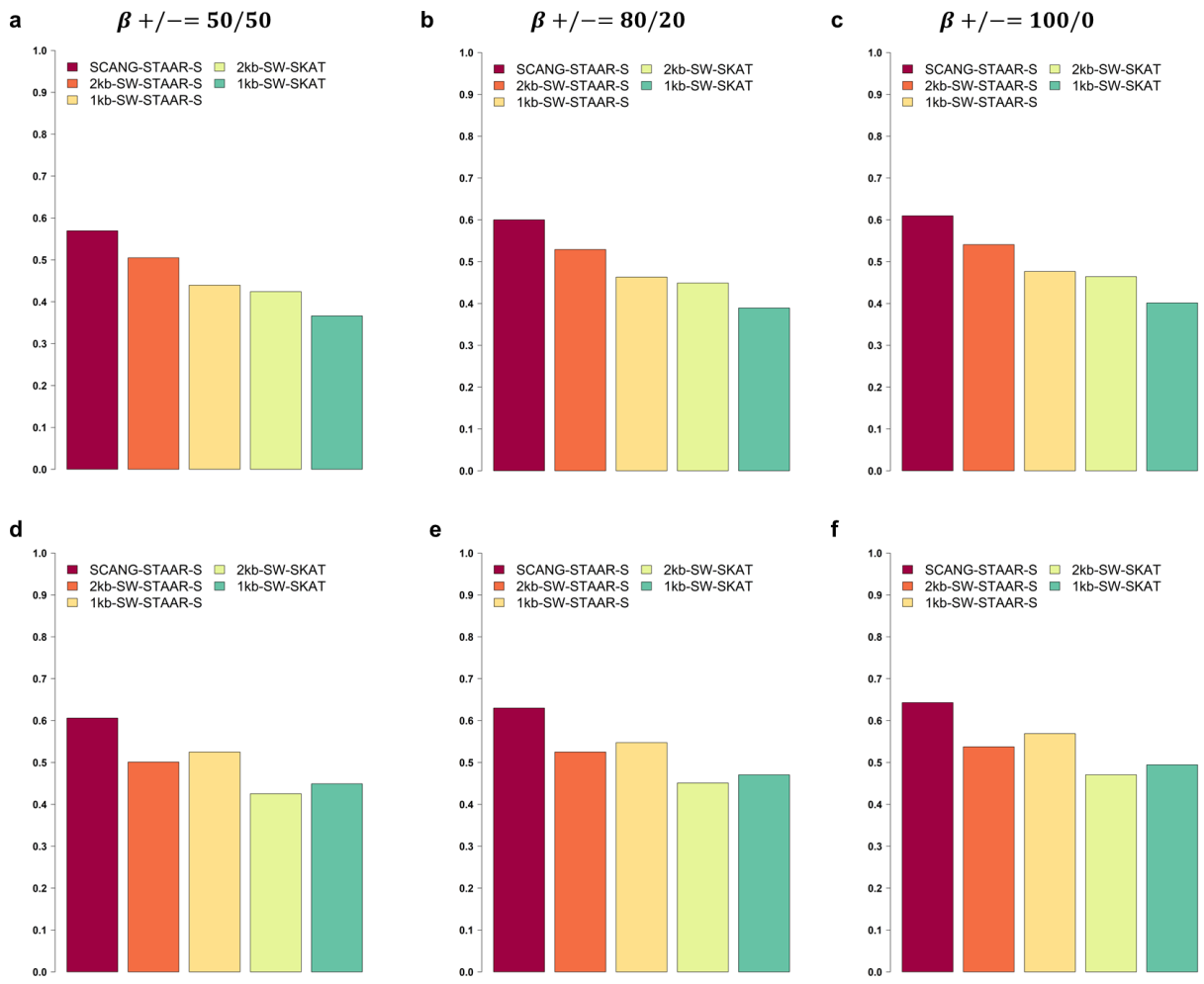
Supplementary Figure 13. Power comparisons of dynamic window procedure SCANG-STAAR-B and the corresponding sliding window procedures using burden for continuous trait analysis. Empirical power was evaluated by the causal variants detection rate and the signal region detection rate defined in the simulation section. Both criteria were calculated at the genome-wise/family-wise type I error $\alpha = 0.01$. We randomly selected two signal regions (variant phenotype association regions) across the 10-Mb genome in each simulation replicate. The length of the signal regions was randomly selected from 1 kb, 1.5 kb, and 2 kb. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model, and on average there were 15% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$, where c_0 was set to be 0.10. From left to right, the plots consider settings in which the effect sizes for the causal variants are 50% positive (50% negative), 80% positive (20% negative) and 100% positive (0% negative). The causal variants detection rate (A-C) is the proportion of detected causal variants in 1,000 simulated whole-genome data sets. A causal variant is defined as detected if it is in one of the detected signal regions. The signal region detection rate (D-F) is the proportion of detected signal regions in 1,000 simulated whole-genome data sets. A signal region is defined as detected if it overlapped with one of the detected signal regions. For each configuration, the sample size was 50,000. For each setting, five methods were compared: SCANG-STAAR-B and 1-kb and 2-kb sliding window procedures (SW for short) using burden and STAAR-B. For SCANG-STAAR-B, the range of search window lengths was set by the number of variants in searching windows between 40 and 300.


a
β +/−= 50/50
b
β +/−= 80/20
c
β +/−= 100/0
d
e
f
SCANG-STAAR-B
2kb-SW-STAAR-B
1kb-SW-STAAR-B
2kb-SW-Burden
1kb-SW-Burden
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

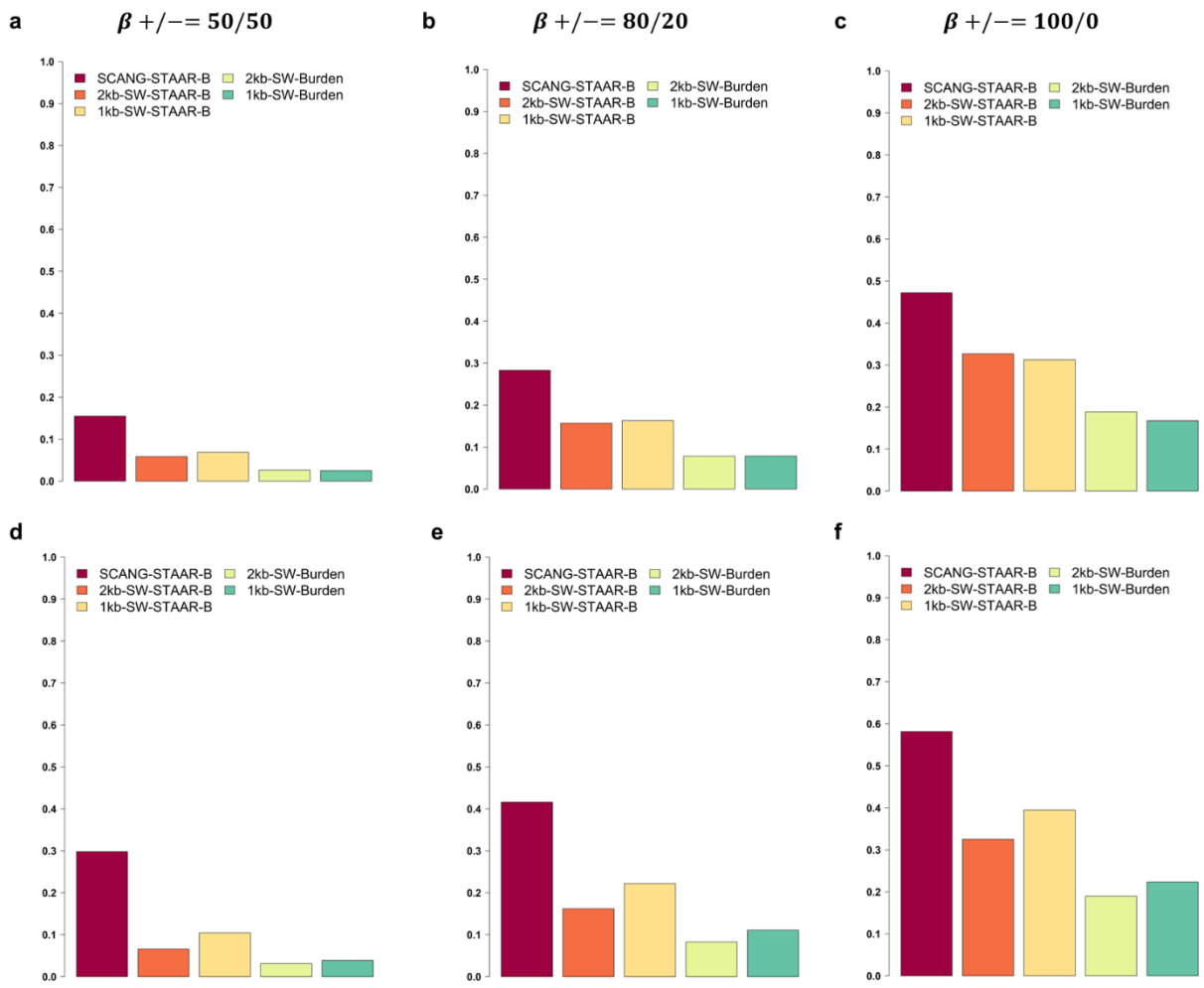
Supplementary Figure 14. Power comparisons of dynamic window procedure SCANG-STAAR and sliding window procedures using burden, SKAT, SKAT-O, and STAAR for dichotomous trait analysis. Empirical power was evaluated by the causal variants detection rate and the signal region detection rate defined in the simulation section. Both criteria were calculated at the genome-wise/family-wise type I error $\alpha = 0.01$. We randomly selected two signal regions (variant phenotype association regions) across the 10-Mb genome in each simulation replicate. The length of the signal regions was randomly selected from 1 kb, 1.5 kb, and 2 kb. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model, and on average there were 15% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0|\log_{10} MAF_j|$, where c_0 was set to be 0.14 and gives an odds ratio of 2 for a variant with MAF of 1×10^{-5}. From left to right, the plots consider settings in which the effect sizes for the causal variants are 50% positive (50% negative), 80% positive (20% negative) and 100% positive (0% negative). The causal variants detection rate (A-C) is the proportion of detected causal variants in 1,000 simulated whole-genome data sets. A causal variant is defined as detected if it is in one of the detected signal regions. The signal region detection rate (D-F) is the proportion of detected signal regions in 1,000 simulated whole-genome data sets. A signal region is defined as detected if it overlapped with one of the detected signal regions. For each configuration, the sample size was 50,000. For each setting, nine methods were compared: SCANG-STAAR-S and 1-kb and 2-kb sliding window procedures (SW for short) using burden, SKAT, SKAT-O and STAAR-O. For SCANG-STAAR-S, the range of search window lengths was set by the number of variants in searching windows between 40 and 300.


a
β +/−= 50/50
b
β +/−= 80/20
c
β +/−= 100/0
d
e
f
SCANG-STAAR
2kb-SW-STAAR
1kb-SW-STAAR
2kb-SW-SKAT-O
1kb-SW-SKAT-O
2kb-SW-SKAT
1kb-SW-SKAT
2kb-SW-Burden
1kb-SW-Burden
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

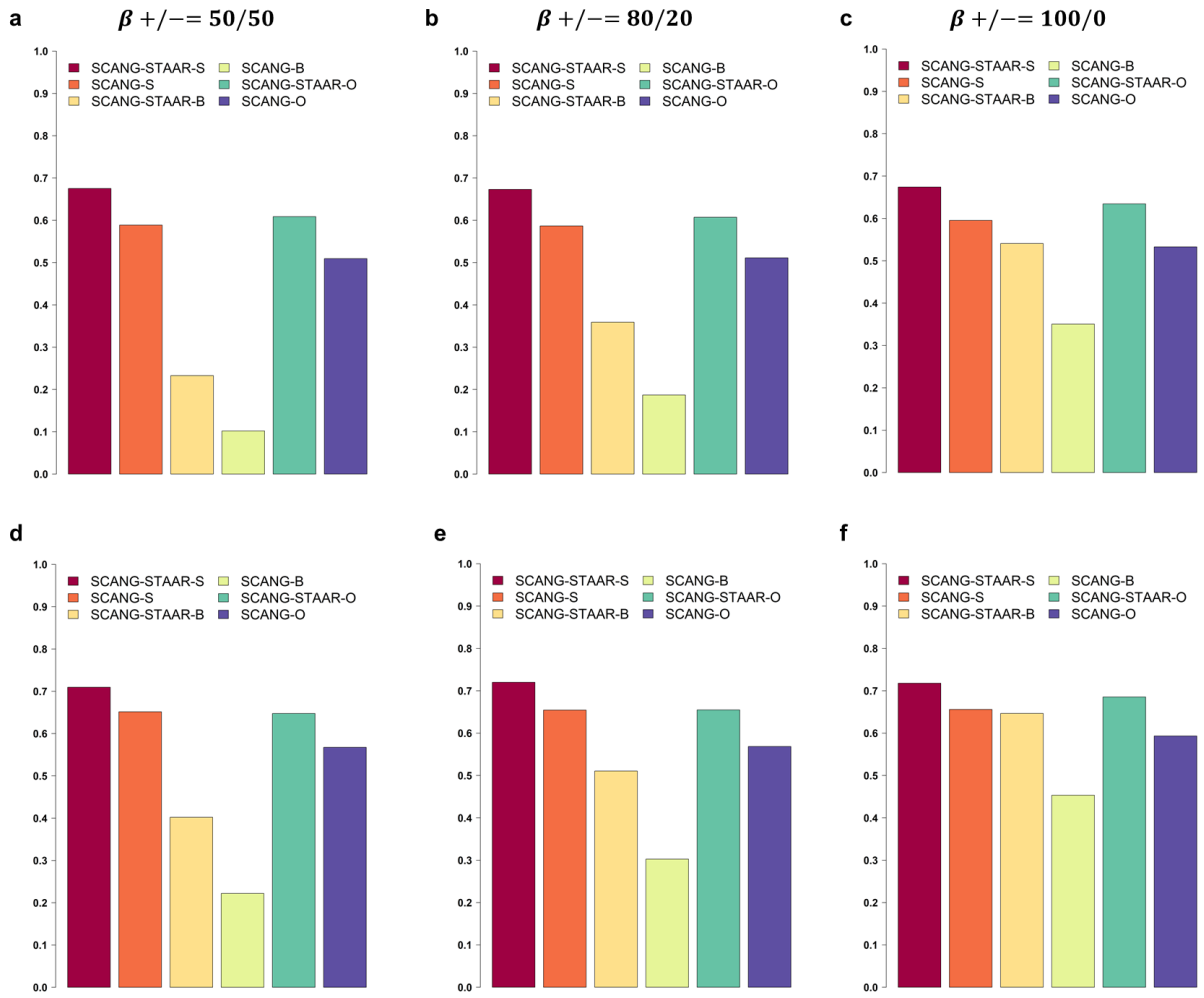
Supplementary Figure 15. Power comparisons of dynamic window procedure SCANG-STAAR-S and the corresponding sliding window procedures using SKAT for dichotomous trait analysis. Empirical power was evaluated by the causal variants detection rate and the signal region detection rate defined in the simulation section. Both criteria were calculated at the genome-wise/family-wise type I error $\alpha = 0.01$. We randomly selected two signal regions (variant phenotype association regions) across the 10-Mb genome in each simulation replicate. The length of the signal regions was randomly selected from 1 kb, 1.5 kb, and 2 kb. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model, and on average there were 15% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0|\log_{10} MAF_j|$, where c_0 was set to be 0.14 and gives an odds ratio of 2 for a variant with MAF of 1×10^{-5}. From left to right, the plots consider settings in which the effect sizes for the causal variants are 50% positive (50% negative), 80% positive (20% negative) and 100% positive (0% negative). The causal variants detection rate (A-C) is the proportion of detected causal variants in 1,000 simulated whole-genome data sets. A causal variant is defined as detected if it is in one of the detected signal regions. The signal region detection rate (D-F) is the proportion of detected signal regions in 1,000 simulated whole-genome data sets. A signal region is defined as detected if it overlapped with one of the detected signal regions. For each configuration, the sample size was 50,000. For each setting, five methods were compared: SCANG-STAAR-S and 1-kb and 2-kb sliding window procedures (SW for short) using SKAT and STAAR-S. For SCANG-STAAR-S, the range of search window lengths was set by the number of variants in searching windows between 40 and 300.


a
β +/−= 50/50
b
β +/−= 80/20
c
β +/−= 100/0
d
e
f
SCANG-STAAR-S
2kb-SW-STAAR-S
1kb-SW-STAAR-S
2kb-SW-SKAT
1kb-SW-SKAT
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

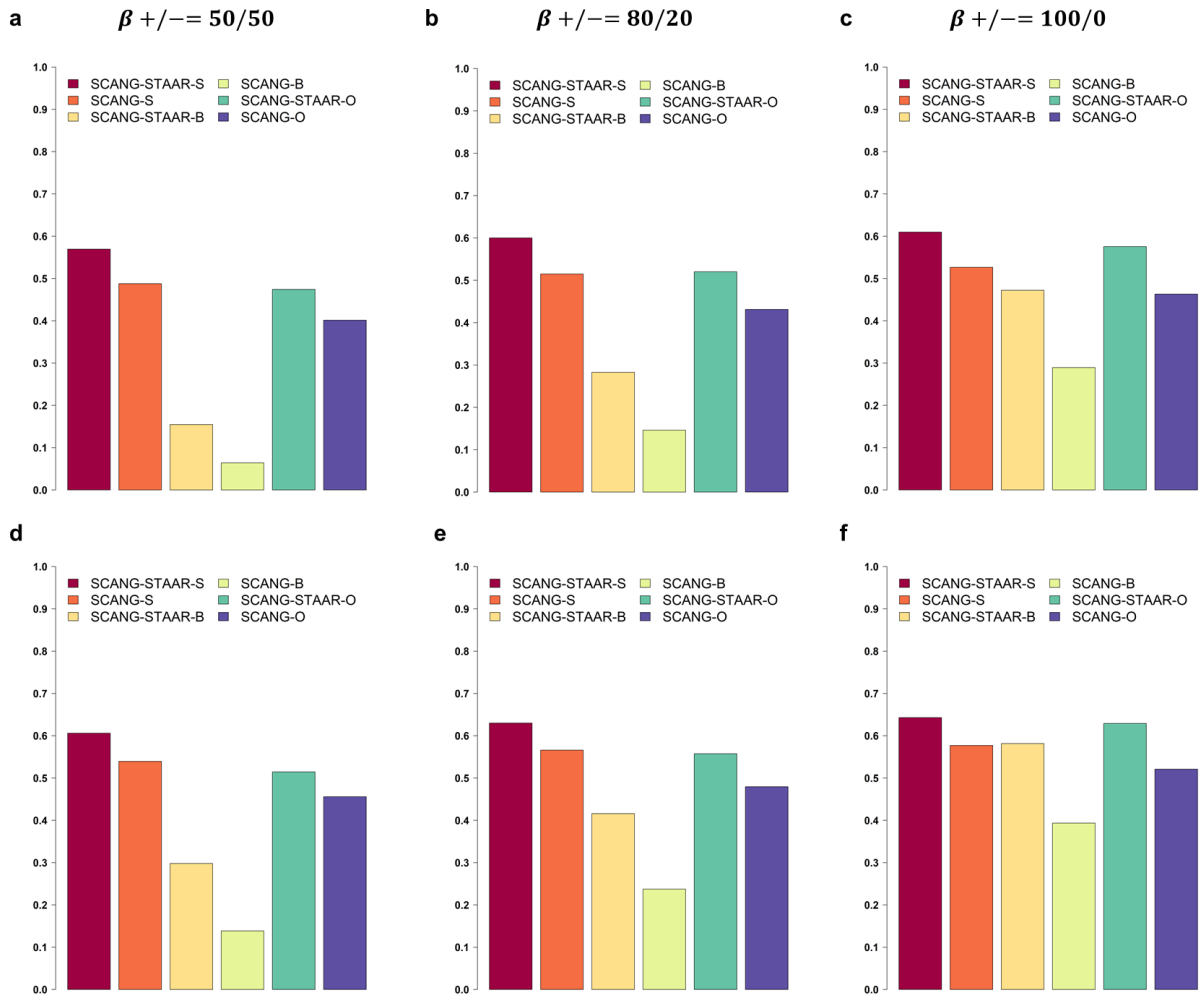
Supplementary Figure 16. Power comparisons of dynamic window procedure SCANG-STAAR-B and the corresponding sliding window procedures using burden for dichotomous trait analysis. Empirical power was evaluated by the causal variants detection rate and the signal region detection rate defined in the simulation section. Both criteria were calculated at the genome-wise/family-wise type I error $\alpha = 0.01$. We randomly selected two signal regions (variant phenotype association regions) across the 10-Mb genome in each simulation replicate. The length of the signal regions was randomly selected from 1 kb, 1.5 kb, and 2 kb. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model, and on average there were 15% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0|\log_{10} MAF_j|$, where c_0 was set to be 0.14 and gives an odds ratio of 2 for a variant with MAF of 1×10^{-5}. From left to right, the plots consider settings in which the effect sizes for the causal variants are 50% positive (50% negative), 80% positive (20% negative) and 100% positive (0% negative). The causal variants detection rate (A-C) is the proportion of detected causal variants in 1,000 simulated whole-genome data sets. A causal variant is defined as detected if it is in one of the detected signal regions. The signal region detection rate (D-F) is the proportion of detected signal regions in 1,000 simulated whole-genome data sets. A signal region is defined as detected if it overlapped with one of the detected signal regions. For each configuration, the sample size was 50,000. For each setting, five methods were compared: SCANG-STAAR-B and 1-kb and 2-kb sliding window procedures (SW for short) using burden and STAAR-B. For SCANG-STAAR-B, the range of search window lengths was set by the number of variants in searching windows between 40 and 300.


a
β +/−= 50/50
b
β +/−= 80/20
c
β +/−= 100/0
d
e
f
SCANG-STAAR-B
2kb-SW-STAAR-B
1kb-SW-STAAR-B
2kb-SW-Burden
1kb-SW-Burden
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

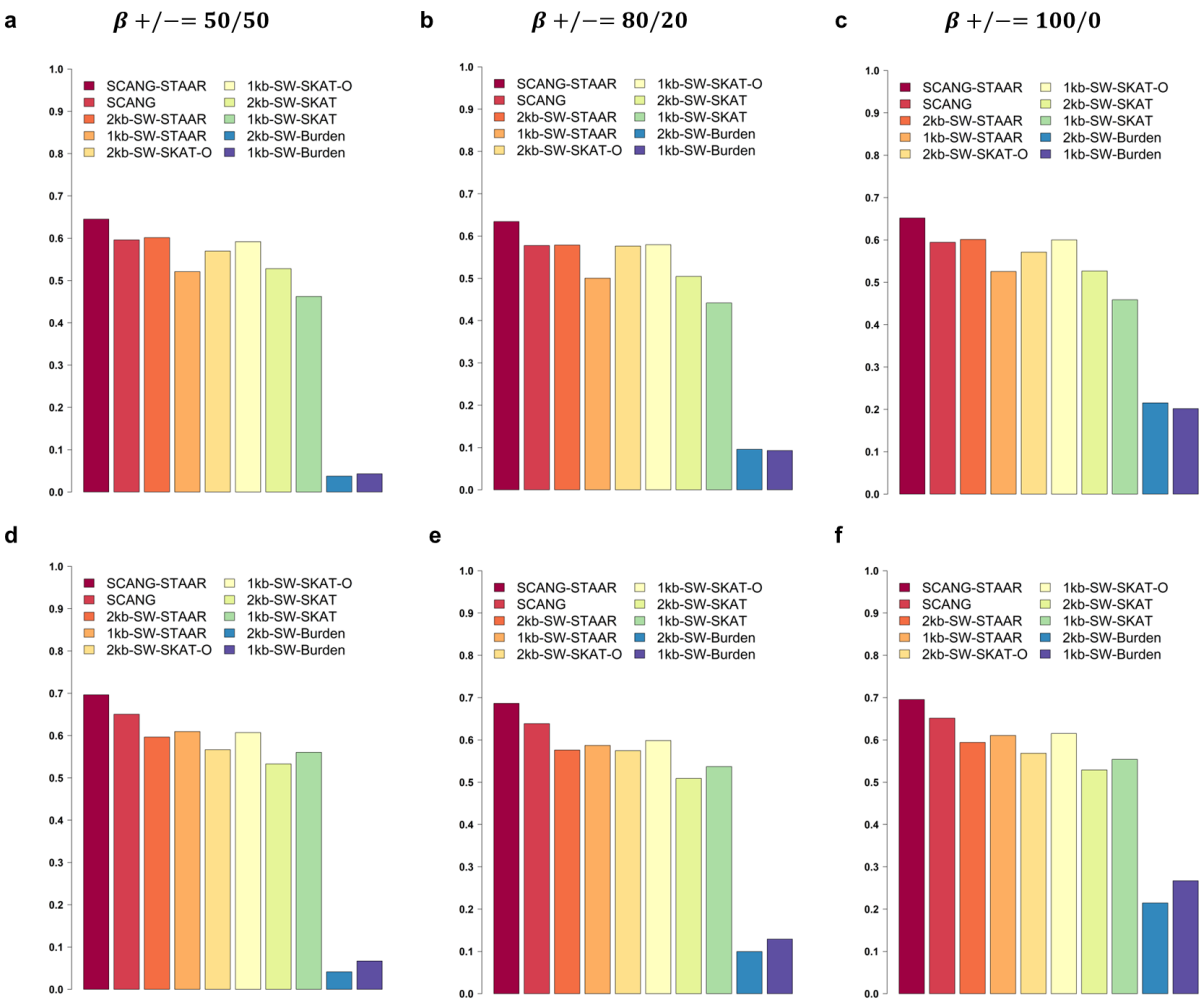
Supplementary Figure 17. Power comparisons of dynamic window procedure SCANG-STAAR and SCANG for continuous trait analysis. Empirical power was evaluated by the causal variants detection rate and the signal region detection rate defined in the simulation section. Both criteria were calculated at the genome-wise/family-wise type I error $\alpha = 0.01$. We randomly selected two signal regions (variant phenotype association regions) across the 10-Mb genome in each simulation replicate. The length of the signal regions was randomly selected from 1 kb, 1.5 kb, and 2 kb. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model, and on average there were 15% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$, where c_0 was set to be 0.14 and gives an odds ratio of 2 for a variant with MAF of 1×10^{-5}. From left to right, the plots consider settings in which the effect sizes for the causal variants are 50% positive (50% negative), 80% positive (20% negative) and 100% positive (0% negative). The causal variants detection rate (A-C) is the proportion of detected causal variants in 1,000 simulated whole-genome data sets. A causal variant is defined as detected if it is in one of the detected signal regions. The signal region detection rate (D-F) is the proportion of detected signal regions in 1,000 simulated whole-genome data sets. A signal region is defined as detected if it is overlapped with one of the detected signal regions. For each configuration, the sample size was 50,000. For each setting, six methods were compared: SCANG-STAAR-S, SCANG-S, SCANG-STAAR-B, SCANG-B, SCANG-STAAR-O, SCANG-O. For all six methods, the range of search window lengths was set by the number of variants in searching windows between 40 and 300.


a
β +/−= 50/50
b
β +/−= 80/20
c
β +/−= 100/0
d
e
f
SCANG-STAAR-S
SCANG-S
SCANG-STAAR-B
SCANG-B
SCANG-STAAR-O
SCANG-O
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

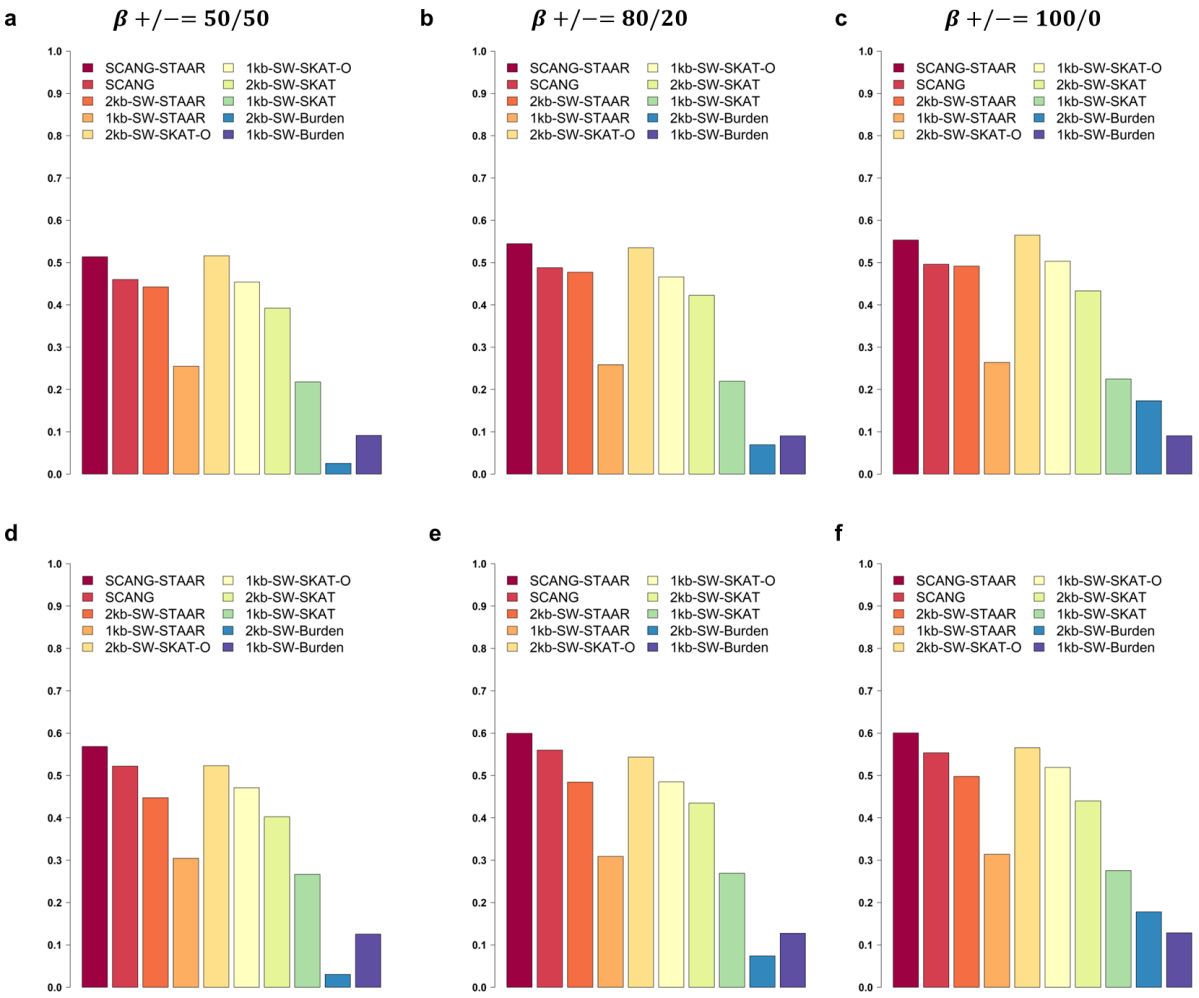
Supplementary Figure 18. Power comparisons of dynamic window procedure SCANG-STAAR and SCANG for dichotomous trait analysis. Empirical power was evaluated by the causal variants detection rate and the signal region detection rate defined in the simulation section. Both criteria were calculated at the genome-wise/family-wise type I error $\alpha = 0.01$. We randomly selected two signal regions (variant phenotype association regions) across the 10-Mb genome in each simulation replicate. The length of the signal regions was randomly selected from 1 kb, 1.5 kb, and 2 kb. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model, and on average there were 15% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$, where c_0 was set to be 0.14 and gives an odds ratio of 2 for a variant with MAF of 1×10^{-5}. From left to right, the plots consider settings in which the effect sizes for the causal variants are 50% positive (50% negative), 80% positive (20% negative) and 100% positive (0% negative). The causal variants detection rate (A-C) is the proportion of detected causal variants in 1,000 simulated whole-genome data sets. A causal variant is defined as detected if it is in one of the detected signal regions. The signal region detection rate (D-F) is the proportion of detected signal regions in 1,000 simulated whole-genome data sets. A signal region is defined as detected if it overlapped with one of the detected signal regions. For each configuration, the sample size was 50,000. For each setting, six methods were compared: SCANG-STAAR-S, SCANG-S, SCANG-STAAR-B, SCANG-B, SCANG-STAAR-O, SCANG-O. For all six methods, the range of search window lengths was set by the number of variants in searching windows between 40 and 300.


a
β +/−= 50/50
b
β +/−= 80/20
c
β +/−= 100/0
d
e
f
SCANG-STAAR-S
SCANG-S
SCANG-STAAR-B
SCANG-B
SCANG-STAAR-O
SCANG-O
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Supplementary Figure 19. Power comparisons of dynamic window procedure SCANG-STAAR, SCANG and sliding window procedures using burden, SKAT, SKAT-O, and STAAR for continuous trait analysis where no annotations were informative. Empirical power was evaluated by the causal variants detection rate and the signal region detection rate defined in the simulation section. Both criteria were calculated at the genome-wise/family-wise type I error $\alpha = 0.01$. We randomly selected two signal regions (variant phenotype association regions) across the 10-Mb genome in each simulation replicate. The length of the signal regions was randomly selected from 1 kb, 1.5 kb, and 2 kb. 15% of variants within each signal region were randomly chosen as causal variants without using the annotation information. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$, where c_0 was set to be 0.10. From left to right, the plots consider settings in which the effect sizes for the causal variants are 50% positive (50% negative), 80% positive (20% negative) and 100% positive (0% negative). The causal variants detection rate (A-C) is the proportion of detected causal variants in 1,000 simulated whole-genome data sets. A causal variant is defined as detected if it is in one of the detected signal regions. The signal region detection rate (D-F) is the proportion of detected signal regions in 1,000 simulated whole-genome data sets. A signal region is defined as detected if it overlapped with one of the detected signal regions. For each configuration, the sample size was 50,000. For each setting, ten methods were compared: SCANG-STAAR-S, SCANG-S and 1-kb and 2-kb sliding window procedures (SW for short) using burden, SKAT, SKAT-O and STAAR-O. For SCANG-STAAR-S and SCANG-S, the range of search window lengths was set by the number of variants in searching windows between 40 and 300.


a
β +/−= 50/50
b
β +/−= 80/20
c
β +/−= 100/0
d
e
f
SCANG-STAAR
SCANG
2kb-SW-STAAR
1kb-SW-STAAR
2kb-SW-SKAT-O
1kb-SW-SKAT-O
2kb-SW-SKAT
1kb-SW-SKAT
2kb-SW-Burden
1kb-SW-Burden
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Supplementary Figure 20. Power comparisons of dynamic window procedure SCANG-STAAR, SCANG and sliding window procedures using burden, SKAT, SKAT-O, and STAAR for dichotomous trait analysis. Empirical power was evaluated by the causal variants detection rate and the signal region detection rate defined in the simulation section. Both criteria were calculated at the genome-wise/family-wise type I error $\alpha = 0.01$. We randomly selected two signal regions (variant phenotype association regions) across the 10-Mb genome in each simulation replicate. The length of the signal regions was randomly selected from 1 kb, 1.5 kb, and 2 kb. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model, and on average there were 15% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0|\log_{10} MAF_j|$, where c_0 was set to be 0.14 and gives an odds ratio of 2 for a variant with MAF of 1×10^{-5}. From left to right, the plots consider settings in which the effect sizes for the causal variants are 50% positive (50% negative), 80% positive (20% negative) and 100% positive (0% negative). The causal variants detection rate (A-C) is the proportion of detected causal variants in 1,000 simulated whole-genome data sets. A causal variant is defined as detected if it is in one of the detected signal regions. The signal region detection rate (D-F) is the proportion of detected signal regions in 1,000 simulated whole-genome data sets. A signal region is defined as detected if it overlapped with one of the detected signal regions. For each configuration, the sample size was 50,000. For each setting, nine methods were compared: SCANG-STAAR-S, SCANG-S and 1-kb and 2-kb sliding window procedures (SW for short) using burden, SKAT, SKAT-O and STAAR-O. For SCANG-STAAR-S and SCANG-S, the range of search window lengths was set by the number of variants in searching windows between 40 and 300.


a
β +/−= 50/50
b
β +/−= 80/20
c
β +/−= 100/0
d
e
f
SCANG-STAAR
SCANG
2kb-SW-STAAR
1kb-SW-STAAR
2kb-SW-SKAT-O
1kb-SW-SKAT-O
2kb-SW-SKAT
1kb-SW-SKAT
2kb-SW-Burden
1kb-SW-Burden
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

Supplementary Note

TOPMed study participants and acknowledgements

Discovery phase (n = 21,015)

Framingham Heart Study (FHS)

The FHS is a three generational prospective cohort that has been described in detail previously[1]. Individuals were initially recruited in 1948 in Framingham, USA to evaluate cardiovascular disease risk factors. The second generation cohort (5,124 offspring of the original cohort) was recruited between 1971 and 1975[2, 3]. The third generation cohort (4,095 grandchildren of the original cohort) was collected between 2002 and 2005. Fasting lipid levels were measured at exam 1 of the Offspring (1971-1975) and third generation (2002-2005) cohorts, using standard LRC protocols.

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) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C).

The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195, HHSN268201500001I and 75N92019D00031 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. Dr. Vasan is supported in part by the Evans Medical Foundation and the Jay and Louis Coffman Endowment from the Department of Medicine, Boston University School of Medicine.

Jackson Heart Study (JHS)

The JHS is a large, population-based observational study evaluating the etiology of cardiovascular, renal, and respiratory diseases among African Americans residing in the

three counties (Hinds, Madison, and Rankin) that make up the Jackson, Mississippi metropolitan area[4, 5]. Data and biologic materials have been collected from 5,306 participants, including a nested family cohort of 1,498 members of 264 families. The age at enrollment for the unrelated cohort was 35-84 years; the family cohort included related individuals >21 years old. Participants provided an extensive medical and social history and had an array of physical and biochemical measurements and diagnostic procedures, and a subset of participants provided genomic DNA during a baseline examination (2000-2004) and two follow-up examinations (2005-2008 and 2009-2012), with a fourth examination ongoing. Annual follow-up interviews and cohort surveillance are ongoing.

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: The Jackson Heart Study” (phs000964.v1.p1) was performed at the University of Washington Northwest Genomics Center (HHSN268201100037C).

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I/HHSN26800001) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) 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.

Multi-Ethnic Study of Atherosclerosis (MESA)

The Multi-Ethnic Study of Atherosclerosis is a National Heart, Lung and Blood Institute-sponsored, population-based investigation of subclinical cardiovascular disease and its progression[6]. A total of 6,814 individuals, aged 45 to 84 years, were recruited from six US communities (Baltimore City and County, MD; Chicago, IL; Forsyth County, NC; Los

Angeles County, CA; New York, NY; and St. Paul, MN) between July 2000 and August 2002. Participants were excluded if they had physician-diagnosed cardiovascular disease prior to enrollment, including angina, myocardial infarction, heart failure, stroke or TIA, resuscitated cardiac arrest or a cardiovascular intervention (e.g., CABG, angioplasty, valve replacement, or pacemaker/defibrillator placement). Pre-specified recruitment plans identified four racial/ethnic groups (White European-American, African-American, Hispanic-American, and Chinese-American) for enrollment, with targeted oversampling of minority groups to enhance statistical power.

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: Multi-Ethnic Study of Atherosclerosis (MESA)” (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN268201800001I).

The MESA project is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

MESA Family is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL071259, by the National Center for Research Resources, Grant UL1RR033176. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

Old Order Amish (OOA)

The Amish Complex Disease Research Program includes a set of large community-based studies focused largely on cardiometabolic health carried out in the Old Order Amish (OOA) community of Lancaster, Pennsylvania[7]. The OOA population of Lancaster County, PA immigrated to the Colonies from Western Europe in the early 1700's. There are now over 38,000 OOA individuals in the Lancaster area, nearly all of whom can trace their ancestry back 12-14 generations to approximately 400 founders. Investigators at the University of Maryland School of Medicine have been studying the genetic determinants of cardiometabolic health in this population since 1993. To date, over 8,000 Amish adults have participated in one or more of our studies.
The 1,123 Amish subjects included in the TOPMed program were enrolled in studies supported by NIH grants R01 AG18728, U01 HL072515, R01 HL088119, R01 HL121007, and P30 DK072488. WGS for “NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish” (phs000956) was performed at the Broad Institute of MIT and Harvard (3R01HL121007-01S1).

Genome-wide Association Study of Adiposity in Samoans (Samoan)

The parent Samoan study is a population-based genome-wide association study (GWAS) of adiposity and cardiometabolic phenotypes among adults, 25-65 years of age, from the independent nation of Samoa in the South Pacific. The research goal of this study is to identify genetic variation that increases susceptibility to obesity and

cardiometabolic phenotypes. Biomarker and questionnaire data were collected to assess cardiometabolic phenotypes. DNA was collected and the Affymetrix 6.0 chip used for SNP genotyping. After quality control checks on genotyping and excluding individuals with key missing data we have a final sample of 3,122 adults with high-quality genome-wide marker data[8]. Participation in TOPMed provided whole genome sequence data for 1,285 individuals from the GWAS sample chosen for maximal informativity for our Samoan-specific imputation panel.

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: Genome-wide Association Study of Adiposity in Samoans" (phs000972) was performed at the University of Washington Northwest Genomics Center (HHSN268201100037C) and the New York Genome Center (HHSN268201500016C).

Data collection was funded by NIH grant R01-HL093093 and R01-HL133040. We thank the Samoan participants of the study and local village authorities. We acknowledge the support of the Samoan Ministry of Health and the Samoa Bureau of Statistics for their support of this research.

Women's Health Initiative (WHI)

The Women's Health Initiative (WHI) is a large study of postmenopausal women's health investigating risk factors for cancer, CVD, age-related fractures and chronic disease[9]. It began in 1993 as a set of randomized controlled clinical trials (CT) and an observational study (OS). Specifically, the CT (n=68,132) included three overlapping components: The Hormone Therapy (HT) Trials (n=27,347), Dietary Modification (DM) Trial (n=48,835), and Calcium and Vitamin D (CaD) Trial (n=36,282). Eligible women could be randomized into as many as all three CTs components. Women who were ineligible or unwilling to join the CT were then invited to join the OS (n=93,676).

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: Women’s Health Initiative” (phs001237) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C).

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 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Long%20List.pdf.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Replication phase (n = 9,123)

Atherosclerosis Risk in Communities Study (ARIC)

The ARIC study is a population-based prospective cohort study of cardiovascular disease sponsored by the National Heart, Lung, and Blood Institute (NHLBI). ARIC included 15,792 individuals, predominantly European American and African American, aged 45-64 years at baseline (1987-89), chosen by probability sampling from four US communities. Cohort members completed three additional triennial follow-up examinations, a fifth exam in 2011-2013, a sixth exam in 2016-2017, a seventh exam in 2018-2019, and an eighth exam in 2020. The ARIC study has been described in detail previously[10].

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: Atherosclerosis Risk in Communities (ARIC)”

(phs001211) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201500015C and 3U54HG003273-12S2) and the Broad Institute of MIT and Harvard (3R01HL092577-06S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN2682018000021). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN2682017000011, HHSN2682017000021, HHSN2682017000031, HHSN2682017000041 and HHSN2682017000051). The authors thank the staff and participants of the ARIC study for their important contributions.

Cleveland Family Study (CFS)

The CFS is a family-based longitudinal study that includes participants with laboratory diagnosed sleep apnea, their family members and neighborhood control families followed between 1990 and 2006. Four examinations over 16 years provided measurements of sleep apnea with overnight polysomnography, anthropometry, and other related phenotypes, as detailed previously[7, 11]. After an overnight fast, blood was collected which was assayed for lipid levels at the University of Vermont Laboratory for Clinical Biochemistry Research. Lipids (triglycerides, HDL cholesterol) from fasted blood serum were measured by enzymatic methods using Centers for Disease Control and Prevention guidelines[12].

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: Cleveland Family Study” (phs000954) was performed at the University of Washington Northwest Genomics Center (3R01HL098433-05S1).

This research was supported by grants HL 046389; HL113338;1R35HL135818 from the National Heart, Lung, and Blood Institute (NHLBI).

Cardiovascular Health Study (CHS)

The Cardiovascular Health Study is a prospective population-based cohort study of risk factors for CHD and stroke in adults 65 years and older[13]. The main objective is to identify factors related to the onset and course of heart disease and stroke. The four Field Centers are located in Forsyth County, NC; Sacramento County, CA; Washington County, MD; and Pittsburgh, PA. The original cohort of 5201 elderly were recruited in 1989-1990; and in 1992-1993, 687 additional minority participants were recruited and examined. Each community sample was obtained from random samples of the Medicare eligibility lists of the Health Care Financing Administration (HCFA). Eligible to participate were persons living in the household of each sampled individual who were: 1) 65 yr or older; 2) non-institutionalized; 3) expected to remain in the area for 3 yr; and 4) able to give informed consent. Excluded were those wheelchair-bound, receiving hospice care or cancer treatment. The minority cohort was recruited using similar methods. Participants were eligible whether or not they had clinically apparent cardiovascular disease. Subjects were followed with semi-annual contacts, alternating between telephone calls and surveillance clinic visits.

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: Cardiovascular Health Study” (phs001368) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201500015C).

This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006, and grants U01HL080295 and U01HL130114 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org.

Diabetes Heart Study (DHS)

The Diabetes Heart Study (DHS) began as a family-based study enriched for type 2 diabetes (T2D). The initial cohort included 1443 European American and African American participants from 564 families with multiple cases of type 2 diabetes recruited between 1998 and 2006[14]. As an ancillary study, the African American Diabetes Heart Study (AA-DHS) expanded the total number of African Americans to 691 by recruiting additional unrelated participants with type 2 diabetes from 2007 and 2010[15]. All participants were extensively phenotyped for measures of subclinical CVD and other known CVD risk factors. Primary outcomes were quantified burden of vascular calcified plaque in the coronary artery, carotid artery, and abdominal aorta all determined from non-contrast computed tomography scans. For TOPMed, DHS and AA-DHS African American participants with CAC were selected for WGS, prioritizing the inclusion of families.

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: Diabetes Heart Study” (phs001412) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C).

This work was supported by R01 HL92301, R01 HL67348, R01 NS058700, R01 AR48797, R01 DK071891, R01 AG058921, the General Clinical Research Center of the Wake Forest University School of Medicine (M01 RR07122, F32 HL085989), the

American Diabetes Association, and a pilot grant from the Claude Pepper Older Americans Independence Center of Wake Forest University Health Sciences (P60 AG10484).

Genetic Study of Atherosclerosis Risk (GeneSTAR)

GeneSTAR is an ongoing family-based prospective study designed to determine environmental, phenotypic, and genetic causes of premature cardiovascular disease. GeneSTAR was originally conducted in healthy adult European- and African-American siblings of probands with documented early onset coronary disease under 60 years of age at the time of hospitalization in any of 10 Baltimore area hospitals from 1982-2006. Participants were screened for traditional coronary disease and stroke risk factors and have been followed regularly to ascertain incident cardiovascular disease[16]. Commencing in 2003, the siblings, their offspring, and the coparent of the offspring who were free of cardiovascular disease participated in a 2 week trial of aspirin 81 mg/day with pre and post ex vivo platelet function assessed using multiple agonists and were screened for traditional coronary disease and stroke risk factors[17]. Of the total 3949 participants, 1786 were selected for TOPMed prioritized on complete platelet function measures and largest family size.

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: Genetic Study of Atherosclerosis Risk” (phs001218) was performed at Psomagen (formerly Macrogen; 3R01HL112064-04S1), Illumina (R01HL112064), and the Broad Institute of MIT and Harvard (HHSN268201500014C).

GeneSTAR was supported by grants from the National Institutes of Health/National Heart, Lung, and Blood Institute (U01 HL72518, HL087698, HL49762, HL59684, HL58625, HL071025, HL112064), by a grant from the National Institutes of Health/National Institute of Nursing Research (NR0224103), and by a grant from the

National Institutes of Health/National Center for Research Resources (M01-RR000052) to the Johns Hopkins General Clinical Research Center.

Genetic Epidemiology Network of Arteriopathy (GENOA)

The Genetic Epidemiology Network of Arteriopathy (GENOA) is one of four networks in the NHLBI Family-Blood Pressure Program (FBPP)[18]. GENOA's long-term objective is to elucidate the genetics of target organ complications of hypertension, including both atherosclerotic and arteriolosclerotic complications involving the heart, brain, kidneys, and peripheral arteries[19]. The longitudinal GENOA Study recruited European-American and African-American sibships with at least 2 individuals with clinically diagnosed essential hypertension before age 60 years. All other members of the sibship were invited to participate regardless of their hypertension status. Participants were diagnosed with hypertension if they had either 1) a previous clinical diagnosis of hypertension by a physician with current anti-hypertensive treatment, or 2) an average systolic blood pressure ≥ 140 mm Hg or diastolic blood pressure ≥ 90 mm Hg based on the second and third readings at the time of their clinic visit. Only participants of the African-American Cohort were sequenced through TOPMed.

During the first exam (Phase 1; 1996-2000), 1,583 European-Americans from Rochester, MN and 1,854 African-Americans from Jackson, MS were examined. Between 2000 and 2004 (Phase 2), 1,241 participants of the European-American Cohort and 1,482 participants of the African-American cohort returned for a second examination. The second examination of the European-American cohort included computed tomography scans for coronary artery calcification while the second examination of the African-American cohort included an echocardiogram. Between 2009 and 2011, an examination that included computed tomography scans for coronary artery calcification (CAC Study) was conducted on 752 participants of the African-American Cohort.

Every participant with an echocardiogram was selected for whole genome sequencing (WGS) through TOPMed. We then selected 106 African-American participants who had

a computed tomography scan for coronary artery calcification but not an echocardiogram or were a sibling of someone already selected for WGS. Finally, we excluded individuals whom we knew were already being whole genome sequenced through TOPMed or another sequencing effort (GENOA participants who overlap with ARIC or JHS participants).

Support for GENOA was provided by the National Heart, Lung and Blood Institute (HL054457, HL054464, HL054481, HL119443, HL085571, and HL087660) of the National Institutes of Health. DNA extraction for “NHLBI TOPMed: Genetic Epidemiology Network of Arteriopathy” (phs001345) was performed at the Mayo Clinic Genotyping Core, and WGS was performed at the DNA Sequencing and Gene Analysis Center at the University of Washington (3R01HL055673-18S1) and the Broad Institute (HHSN268201500014C). We would like to thank the GENOA participants.

Genetics of Lipid Lowering Drugs and Diet Network (GOLDN)

GOLDN is a family-based study of European descent individuals recruited in Minneapolis and Salt Lake City (two of the NHLBI Family Heart Study sites). It aims to uncover genetic predictors of variability in lipid phenotypes, which include both fasting and postprandial lipids quantified using traditional methods, NMR, and high-throughput lipidomics. During the initial screening of ~1,350 individuals, the following criteria were used for exclusion: age < 18 years; fasting triglycerides ≥1500 mg/dL; recent history of myocardial infarction, coronary bypass surgery, or coronary angioplasty; self-report of a positive history of liver, kidney, pancreas, or gallbladder disease, or a history of nutrient malabsorption; current use of insulin; abnormal liver or kidney function; in women of childbearing potential, pregnancy, breastfeeding, not using an acceptable form of contraception. Of those who enrolled, 1,048 individuals consented to the use of their DNA in research; 893 participants with data on all exposures, outcomes, and covariates were included in the current study.

GOLDN biospecimens, baseline phenotype data, and intervention phenotype data were collected with funding from National Heart, Lung and Blood Institute (NHLBI) grant U01

HL072524. Whole-genome sequencing in GOLDN was funded by NHLBI grant R01 HL104135-04S1.

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: Genetics of Lipid Lowering Drugs and Diet Network” (phs001359) was performed at the University of Washington Northwest Genomics Center (3R01HL104135-04S1).

San Antonio Family Heart Study (SAFS)

The SAFHS began in 1991, and included 1,431 individuals in 42 extended families at baseline. Probands were 40 to 60 year old low-income Mexican Americans selected at random without regard to presence or absence of disease, almost exclusively from Mexican American census tracts in San Antonio, Texas. All first, second, and third degree relatives of the proband and of the proband's spouse, aged 16 years or above, were eligible to participate in the study. As part of our ongoing studies, we have recruited new family members from the original families, expanding the cohort to almost 3,099 individuals primarily from 73 families. Our study is a mixed longitudinal design. Subjects have been seen between 1 and 4 times with an average of 1.95 examinations.

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: San Antonio Family Heart Study” (phs001215) was performed at the Illumina Genomic Services (3R01HL113323-03S1).

Collection of the San Antonio Family Study data was supported in part by National Institutes of Health (NIH) grants R01 HL045522, MH078143, MH078111 and MH083824; and whole genome sequencing of SAFS subjects was supported by U01 DK085524 and R01 HL113323. We are very grateful to the participants of the San Antonio Family Study for their continued involvement in our research programs.

References

1. Kannel, W.B., Dawber, T.R., Kagan, A., Revotskie, N. & Stokes, J. Factors of risk in the development of coronary heart disease—six-year follow-up experience: the Framingham Study. *Annals of internal medicine* **55**, 33-50 (1961).
2. Kannel, W.B., Feinleib, M., McNamara, P.M., Garrison, R.J. & Castelli, W.P. An investigation of coronary heart disease in families: the Framingham Offspring Study. *American journal of epidemiology* **110**, 281-290 (1979).
3. Splansky, G.L. et al. The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *American journal of epidemiology* **165**, 1328-1335 (2007).
4. Taylor, H.A., Jr. et al. Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. *Ethnicity & disease* **15**, S6-4-17 (2005).
5. Carpenter, M.A. et al. Laboratory, reading center, and coordinating center data management methods in the Jackson Heart Study. *The American journal of the medical sciences* **328**, 131-144 (2004).
6. Bild, D.E. et al. Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology* **156**, 871-881 (2002).
7. Larkin, E.K. et al. A candidate gene study of obstructive sleep apnea in European Americans and African Americans. *American journal of respiratory and critical care medicine* **182**, 947-953 (2010).
8. Minster, R.L. et al. A thrifty variant in CREBRF strongly influences body mass index in Samoans. *Nature genetics* **48**, 1049 (2016).
9. Anderson, G. et al. Design of the Women's Health Initiative clinical trial and observational study. *Controlled clinical trials* **19**, 61-109 (1998).
10. Wright, J.D. et al. The ARIC (atherosclerosis risk in communities) study: JACC focus seminar 3/8. *Journal of the American College of Cardiology* **77**, 2939-2959 (2021).
11. Redline, S., Schluchter, M.D., Larkin, E.K. & Tishler, P.V. Predictors of longitudinal change in sleep-disordered breathing in a nonclinic population. *Sleep* **26**, 703-709 (2003).
12. Cushman, M., Cornell, E.S., Howard, P.R., Bovill, E.G. & Tracy, R.P. Laboratory methods and quality assurance in the Cardiovascular Health Study. *Clinical chemistry* **41**, 264-270 (1995).
13. Fried, L.P. et al. The cardiovascular health study: design and rationale. *Annals of epidemiology* **1**, 263-276 (1991).
14. Bowden, D.W. et al. Review of the Diabetes Heart Study (DHS) family of studies: a comprehensively examined sample for genetic and epidemiological studies of type 2 diabetes and its complications. *The review of diabetic studies: RDS* **7**, 188 (2010).
15. Divers, J. et al. Genome-wide association study of coronary artery calcified atherosclerotic plaque in African Americans with type 2 diabetes. *BMC genetics* **18**, 105 (2017).
16. Vaidya, D. et al. Incidence of coronary artery disease in siblings of patients with premature coronary artery disease: 10 years of follow-up. *The American journal of cardiology* **100**, 1410-1415 (2007).

17. Faraday, N. et al. Relation between atherosclerosis risk factors and aspirin resistance in a primary prevention population. *The American journal of cardiology* **98**, 774-779 (2006).
18. FBPP Investigators Multi-center genetic study of hypertension: the Family Blood Pressure Program (FBPP). *Hypertension* **39**, 3-9 (2002).
19. Daniels, P.R. et al. Familial aggregation of hypertension treatment and control in the Genetic Epidemiology Network of Arteriopathy (GENOA) study. *The American journal of medicine* **116**, 676-681 (2004).

The Samoan Obesity, Lifestyle and Genetic Adaptations Study (OLaGA) Group

Ranjan Deka, Dept. of Environmental Health, University of Cincinnati;

Nicola L. Hawley, Dept. of Chronic Disease Epidemiology, Yale University;

Stephen T. McGarvey, Dept. of Epidemiology and International Health Institute, and Dept. of Anthropology, Brown University;

Ryan L. Minster, Dept. of Human Genetics, University of Pittsburgh;

Take Naseri, Ministry of Health, Government of Samoa;

Muagututi'a Sefuiva Reupena, Lutia I Puava Ae Mapu I Fagalele;

Daniel E. Weeks, Depts. of Human Genetics and Biostatistics, University of Pittsburgh.