SUPPLEMENTAL MATERIAL

Non-conserved lincRNAs associate with complex cardiometabolic disease traits

Andrea S Foulkes, ScD[1,2], Caitlin Selvaggi, MS[1], Tingyi Cao, BS[1,3], Marcella E O'Reilly, PhD[5], Esther Cynn, MS[5], Puyang Ma, BA[4], Heidi Lumish, MD[5], Chenyi Xue, MS[5], Muredach P Reilly, MBBCh, MSCE[5,6]

[1]Biostatistics, Massachusetts General Hospital, Boston, MA 02114; [2]Department of Medicine, Harvard Medical School, Boston, MA 02114. [3]Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA 02115. [4]Data Science, Stanford University, Stanford, CA 94305. [5]Cardiology Division, Department of Medicine and the [6]Irving Institute for Clinical and Translational Sciences, Columbia University, New York, NY 10032.

Major Resources Table

In order to allow validation and replication of experiments, all essential research materials listed in the Methods should be included in the Major Resources Table below. Authors are encouraged to use public repositories for protocols, data, code, and other materials and provide persistent identifiers and/or links to repositories when available. Authors may add or delete rows as needed.

Data & Code Availability

Description	Source / Repository	Persistent ID / URL
Human lincRNAs and protein coding genes	Human GENCODE Release 33	https://www.gencodegenes.org/human/release_33.html
Mouse lincRNAs and protein coding genes	Mouse GENCODE Release 24	https://www.gencodegenes.org/mouse/release_M24.html
Mouse homologs of human genes	Ensembl Biomart	https://www.ensembl.org/biomart/martview/
RepeatMasker data for calculating TE coverage	UCSC	https://genome.ucsc.edu/cgi-bin/hgTables
GWAS summary data for WHRadjBMI and BMI	GIANT & UK BioBank meta-analysis	https://zenodo.org/record/1251813#.X3sbzJNKiHG
GWAS summary data for Height	GIANT & UK BioBank meta-analysis	https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
GWAS summary data for CAD	CARDIoGRAMplusC4D 1000 Genomes-based GWAS	http://www.cardiogramplusc4d.org/data-downloads/
GWAS summary data for T2D	DIAGRAM Consortium T2D GWAS meta-analysis - Adjusted for BMI	https://diagram-consortium.org/downloads.html
GWAS summary data for HDL, LDL, TGs	Global Lipids Genetics Consortium (GLGC) Joint Analysis of Metabochip and GWAS Data	http://csg.sph.umich.edu/willer/public/lipids2013/

Materials and Methods

Pathway analysis. Gene set enrichment tests and functional categorization of the nearest 5' and 3' protein coding genes (PCG) of conserved (243 genes) and non-conserved (84 genes) lincRNAs significantly associated ($P < 5 \times 10^{-8}$) with WHRadjBMI were interrogated separately using Database for Annotation, Visualization and Integrated Discovery Bioinformatics Resources v6.8 (DAVID) (1, 2) in order to characterize the biological pathways associated with each set of lincRNAs. If genes share similar set of terms, they are most likely involved in similar biological mechanisms. The algorithm groups those related genes based on the agreement of sharing similar annotation terms by Kappa statistics(1, 2). Functional annotations of the PCG near conserved and non-conserved WHRadjBMI-associated lincRNAs were analyzed in the context of several databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.ad.jp), Gene Ontology (GO) terms; Biological Processes, Cellular Component, and Functional Annotation; UP_Keywords. Findings were consistent across these different databases and annotations therefore we present targeted results for UniProt Keyword (UP_Keyword) dataset analyses. The cutoff for gene categories was defined at a false discovery rate (FDR) below 0.05.

Supplement Table I. Characteristics of lincRNAs that are not classified based on conservation definitions.

Characteristic[*]	**Absent PCG neighbors (n=1313)**	**Inconsistent PCG rel. orientation (n=88)**
Length	25198 (13900, 66085)	16996 (11840, 27982)
GC content	0.411 (0.379, 0.454)	0.457 (0.415, 0.516)
Exon Count	3 (2, 6)	3 (2, 4)
TE coverage	0.359 (0.176, 0.557)	0.411 (0.201, 0.566)
# SNPs[†]	271 (144, 667)	168 (118, 289)
# SNPs/length[†]	0.010 (0.009, 0.012)	0.010 (0.008, 0.011)

[*] Median and IQR (25th, 75th) across lincRNAs within corresponding category. [†] Summary results for number of SNPs per lincRNA and number of SNPs divided by lincRNA length are based on subset of n=7011 lincRNAs and GWAS SNPs for WHRadjBMI.

Supplement Table II. Complete summary of GWAS signals by cardiometabolic trait and conservation

Conservation defined based on synteny:				
		No signal (n, row %, col %)	**Signal (n, row %, col %)**	**Total**
WHRadjBMI (n=5635)	Non-conserved	1315, 94.3%, 24.6%	80, 5.7%, 27.6%	1395
	Conserved	4030, 95.0%, 75.4%	210, 5.0%, 72.4%	4240
BMI (n=5635)	Non-conserved	1308, 93.8%, 24.9%	87, 6.2%, 23.0%	1395
	Conserved	3949, 93.1%, 75.1%	291, 6.9%, 77.0%	4240
Height (n=5319)	Non-conserved	968, 83.2%, 22.2%	195, 16.8%, 20.0%	1163
	Conserved	3375, 81.2%, 77.8%	781, 18.8%, 80.0%	4156
HDL (n=5395)	Non-conserved	1208, 99.0%, 22.6%	12, 1.0%, 29.3%	1220
	Conserved	4146, 99.3%, 77.45	29, 0.7%, 70.7%	4175
LDL (n=5389)	Non-conserved	1203, 98.8%, 22.5%	14, 1.2%, 36.8%	1217
	Conserved	4148, 99.4%, 77.5%	24, 0.6%, 63.2%	4172
TGs (n=5389)	Non-conserved	1204, 98.9%, 22.5%	13, 1.1%, 29.5%	1217
	Conserved	4141, 99.3%, 77.5%	31, 0.7%, 70.5%	4172
CAD (n=5534)	Non-conserved	1301, 99.4%, 23.6%	8, 0.6%, 33.3%	1309
	Conserved	4209, 99.6%, 76.4%	16, 0.4%, 66.7%	4225
T2D (n=5616)	Non-conserved	1354, 98.4%, 24.4%	22, 1.6%, 31.4%	1376
	Conserved	4192, 98.9%, 75.6%	48, 1.1%, 68.6%	4240
Conservation defined based on synteny and expression:				
		No signal (n, row %, col %)	**Signal (n, row %, col %)**	**Total**
WHRadjBMI (n=5607)	Non-conserved	3173, 94.8%, 59.7%	173, 5.2%, 59.9%	3346
	Conserved	2145, 94.9%, 40.3%	116, 5.1%, 40.1%	2261
BMI (n=5607)	Non-conserved	3115, 93.1%, 59.6%	231, 6.9%, 61.3%	3346
	Conserved	2115, 93.5%, 40.4%	146, 6.5%, 38.7%	2261
Height (n=5292)	Non-conserved	2481, 81.1%, 57.4%	579, 18.9%, 59.6%	3060
	Conserved	1840, 82.4%, 42.6%	392, 17.6%, 40.4%	2232
HDL (n=5368)	Non-conserved	3100, 99.1%, 58.2%	28, 0.9%, 68.3%	3128
	Conserved	2227, 99.4%, 41.8%	13, 0.6%, 31.7%	2240
LDL (n=5362)	Non-conserved	3097, 99.2%, 58.2%	26, 0.8%, 68.4%	3123
	Conserved	2227, 99.5%, 41.8%	12, 0.5%, 31.6%	2239
TGs (n=5362)	Non-conserved	3094, 99.1%, 58.2%	29, 0.9%, 65.9%	3123
	Conserved	2224, 99.3%, 41.8%	15, 0.7%, 34.1%	2239
CAD (n=5506)	Non-conserved	3233, 99.5%, 59.0%	17, 0.5%, 70.8%	3250
	Conserved	2249, 99.7%, 41.0%	7, 0.3%, 29.2%	2256
T2D (n=5588)	Non-conserved	3280, 98.6%, 59.4%	47, 1.4%, 68.1%	3327
	Conserved	2239, 99.0%, 40.6%	22, 1.0%, 31.9%	2261

Supplement Table III. Distribution of GWAS signals for lincRNAs by conservation classification.

		No signal (n, row %, col %)	**Signal (n, row %, col %)**	**Total**
WHRadjBMI (n=5635)	Syntenic or non-syntenic[a]	5345, 94.9%, 79.9%	290, 5.1%, 90.3%	5635
	Absent Neighbor[b]	1269, 98.4%, 19.0%	20, 1.6%, 6.2%	1289
	Inconsistent orientation	76, 87.4%, 1.1%	11, 12.6%, 3.4%	87
BMI (n=5635)	Syntenic or non-syntenic	5257, 93.3%, 80.7%	378, 6.7%, 75.8%	5635
	Absent Neighbor	1178, 91.4%, 18.1%	111, 8.6%, 22.2%	1289
	Inconsistent orientation	77, 88.5%, 1.2%	10, 11.5%, 2.0%	87
Height (n=5319)	Syntenic or non-syntenic	4343, 81.7%, 78.6%	976, 18.3%, 89.8%	5319
	Absent Neighbor	1118, 92.5%, 20.2%	91, 7.5%, 8.4%	1209
	Inconsistent orientation	63, 75.9%, 1.1%	20, 24.1%, 1.8%	83
HDL (n=5395)	Syntenic or non-syntenic	5354, 99.2%, 80.4%	41, 0.8%, 93.2%	5395
	Absent Neighbor	1223, 99.8%, 18.4%	2, 0.2%, 4.5%	1225
	Inconsistent orientation	83, 98.9%, 1.2%	1, 1.2%, 2.3%	84
LDL (n=5389)	Syntenic or non-syntenic	5351, 99.3%, 80.4%	38, 0.7%, 86.4%	5351
	Absent Neighbor	1220, 99.6%, 18.3%	5, 0.4%, 11.4%	1225
	Inconsistent orientation	83, 98.8%, 1.2%	1, 1.2%, 2.3%	84
TGs (n=5389)	Syntenic or non-syntenic	5345, 99.2%, 80.4%	44, 0.8%, 91.7%	5389
	Absent Neighbor	1221, 99.7%, 18.4%	4, 0.3%, 8.3%	1225
	Inconsistent orientation	84, 100.0%, 1.3%	0, 0.0%, 0.0%	84
CAD (n=5534)	Syntenic or non-syntenic	5510, 99.6%, 80.7%	24, 0.4%, 75.0%	5534
	Absent Neighbor	1232, 99.4%, 18.0%	7, 0.6%, 21.9%	1239
	Inconsistent orientation	85, 98.8%, 1.2%	1, 1.2%, 3.1%	86
T2D (n=5616)	Syntenic or non-syntenic	5546, 98.8%, 80.3%	70, 1.2%, 95.9%	5616
	Absent Neighbor	1271, 99.8%, 18.4%	3, 0.2%, 4.1%	1274
	Inconsistent orientation	87, 100.0%, 1.3%	0, 0.0%, 0.0%	87

[a]Syntenic and non-syntenic lincRNAs are considered *classified*. [b]Absent neighbor lincRNAs are considered *unclassified*.

Supplement Table IV. Multivariable adjusted model estimates for effect of classification on GWAS signal by trait

	Estimate for classified*	**Std. Error**	**z value**	**Pr(>|z|)**	**OR (95% CI)**
WHRadjBMI	1.213	0.238	5.091	3.557e-7	3.363 (2.108, 5.364)
BMI	-0.055	0.121	-0.457	0.648	0.946 (0.747, 1.199)
Height	1.055	0.118	8.927	4.383e-19	2.871 (2.277, 3.619)

*Modeling is performed for each trait separately and analysis includes lincRNAs that are classified as syntenic or non-sytenic and lincRNAs that are unclassified, i.e. do not have a neighboring PCG within 900Kb up and/or downstream. LincRNAs with inconsistent orientation between human PCGs and mouse homologs are excluded from this analysis. Analysis is limited to the three traits with GWAS signal in >20 unclassified lincRNAs. P-values correspond to Wald tests of H_0: OR=1 versus the two-sided alternative that the OR is not equal to 1.

Supplement Table V. Results of Database for Annotation, Visualization and Integrated Discovery (DAVID) pathway analysis for protein coding genes (PCGs) near conserved and non-conserved lincRNAs associated with WHRadjBMI: Results for UniProt Keyword (UP_Keyword) annotations are presented.

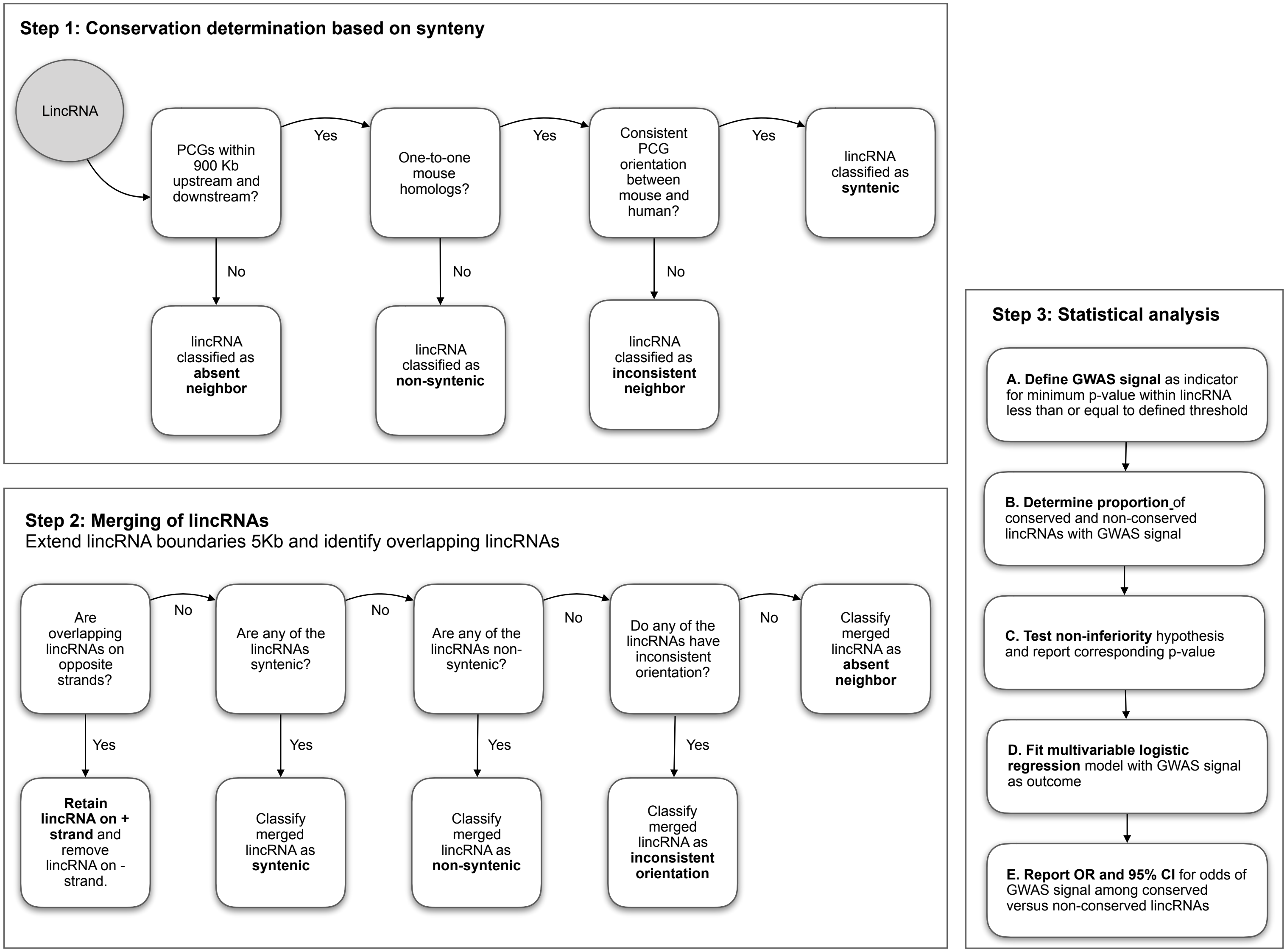
(A) DAVID pathway analysis for PCGs near conserved WHRadjBMI-associated lincRNAs.

Term	Count	%	P-Value	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
DNA-binding	60	24.49	5.53E-11	243	2050	20581	2.4788919	1.47E-08	1.48E-08	1.41E-08
Developmental protein	36	14.69	1.95E-09	243	949	20581	3.21289467	5.18E-07	2.60E-07	2.48E-07
Transcription regulation	57	23.27	1.28E-07	243	2332	20581	2.07017237	3.40E-05	9.00E-06	8.59E-06
Transcription	58	23.67	1.35E-07	243	2398	20581	2.04851437	3.58E-05	9.00E-06	8.59E-06
Nucleus	97	39.59	7.86E-07	243	5244	20581	1.56664014	2.09E-04	4.19E-05	4.01E-05
Homeobox	15	6.122	2.90E-06	243	262	20581	4.84897748	7.72E-04	1.29E-04	1.23E-04
Activator	23	9.388	1.23E-05	243	661	20581	2.9470437	3.27E-03	4.69E-04	4.48E-04
Alternative splicing	155	63.27	7.87E-05	243	10587	20581	1.23999229	2.07E-02	2.63E-03	2.51E-03
Phosphoprotein	126	51.43	1.60E-04	243	8246	20581	1.29415834	4.18E-02	4.76E-03	4.54E-03
Chromosomal rearrangement	13	5.306	6.29E-04	243	334	20581	3.2965304	1.54E-01	1.68E-02	1.60E-02
Disease mutation	48	19.59	0.00111	243	2550	20581	1.59426773	2.57E-01	2.70E-02	2.58E-02
Repressor	17	6.939	0.0018	243	592	20581	2.43213352	3.81E-01	4.01E-02	3.83E-02

(B) DAVID pathway analysis for PCGs near non-conserved WHRadjBMI-associated lincRNAs.

Term	Count	%	P-Value	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
MHC I	7	8.24	7.42E-13	84	10	20581	171.5083333	1.16E-10	1.16E-10	1.1E-10
Immunity	12	14.1	5.20E-06	84	500	20581	5.880285714	8.15E-04	4.08E-04	3.82E-04
Cell division	8	9.41	9.73E-04	84	388	20581	5.051791851	1.42E-01	4.56E-02	4.27E-02
Chromosome	8	9.41	1.16E-03	84	400	20581	4.900238095	1.67E-01	4.56E-02	4.27E-02

Supplement Figure I: Schematic illustration of synteny definition and analytic pipeline.


Step 1: Conservation determination based on synteny
LincRNA
PCGs within 900 Kb upstream and downstream?
Yes
One-to-one mouse homologs?
Yes
Consistent PCG orientation between mouse and human?
Yes
lincRNA classified as **syntenic**
No
lincRNA classified as **absent neighbor**
No
lincRNA classified as **non-syntenic**
No
lincRNA classified as **inconsistent neighbor**
Step 2: Merging of lincRNAs
Extend lincRNA boundaries 5Kb and identify overlapping lincRNAs
Are overlapping lincRNAs on opposite strands?
No
Are any of the lincRNAs syntenic?
No
Are any of the lincRNAs non-syntenic?
No
Do any of the lincRNAs have inconsistent orientation?
No
Classify merged lincRNA as **absent neighbor**
Yes
Retain lincRNA on + strand and remove lincRNA on - strand.
Yes
Classify merged lincRNA as **syntenic**
Yes
Classify merged lincRNA as **non-syntenic**
Yes
Classify merged lincRNA as **inconsistent orientation**
Step 3: Statistical analysis

A. Define GWAS signal as indicator for minimum p-value within lincRNA less than or equal to defined threshold

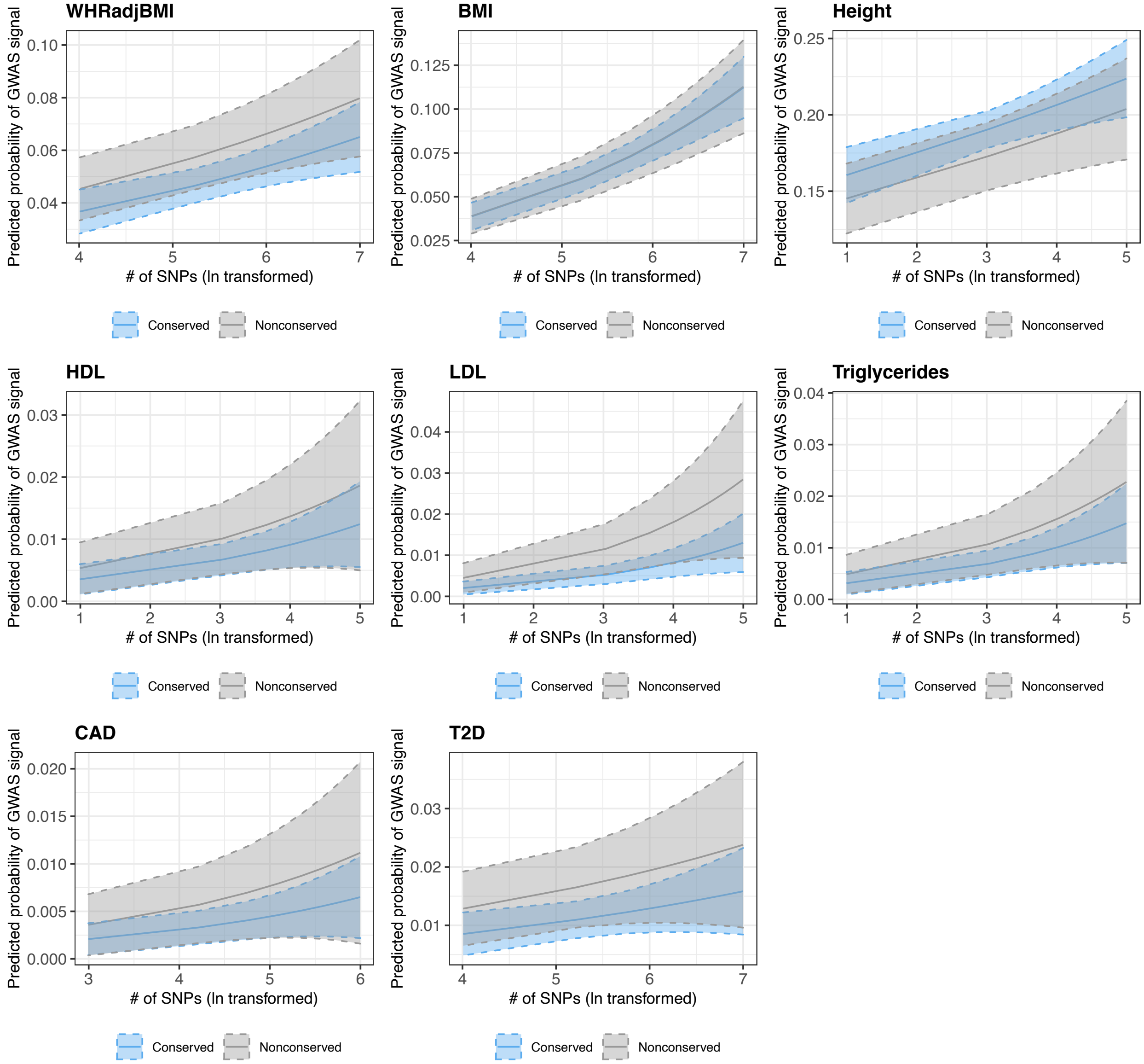
B. Determine proportion of conserved and non-conserved lincRNAs with GWAS signal

C. Test non-inferiority hypothesis and report corresponding p-value

D. Fit multivariable logistic regression model with GWAS signal as outcome

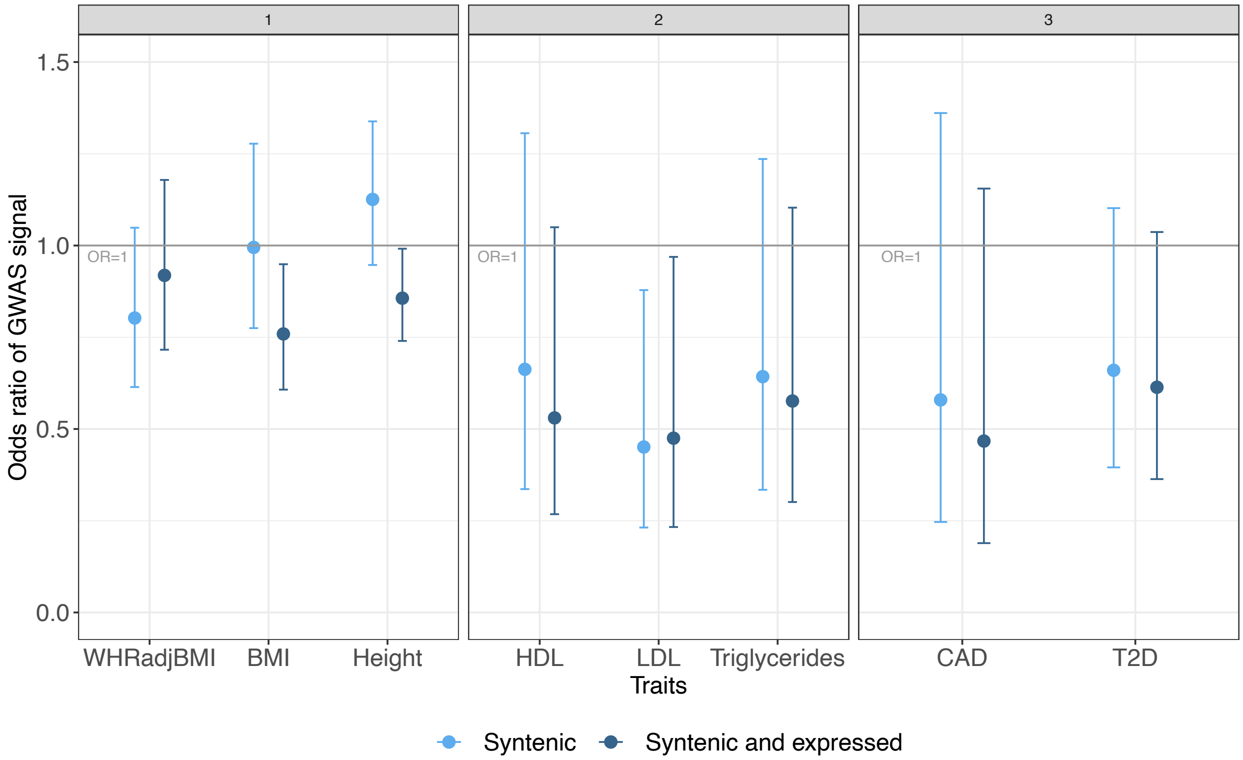
E. Report OR and 95% CI for odds of GWAS signal among conserved versus non-conserved lincRNAs

Supplement Figure II: Predicted probabilities of GWAS signals by number of SNPs and each cardiometabolic trait



Multivariable model-based predictions of the probability of GWAS signals and corresponding prediction intervals are derived separately for each trait (see Table 3a). Median values for all covariates are used as model inputs. As shown, the predicted probability of GWAS signals increases with the number of SNPs and tends to be greater for non-conserved lincRNAs compared to conserved lincRNAs for all traits with the exception of height. Results are based on the primary definition of conservation.

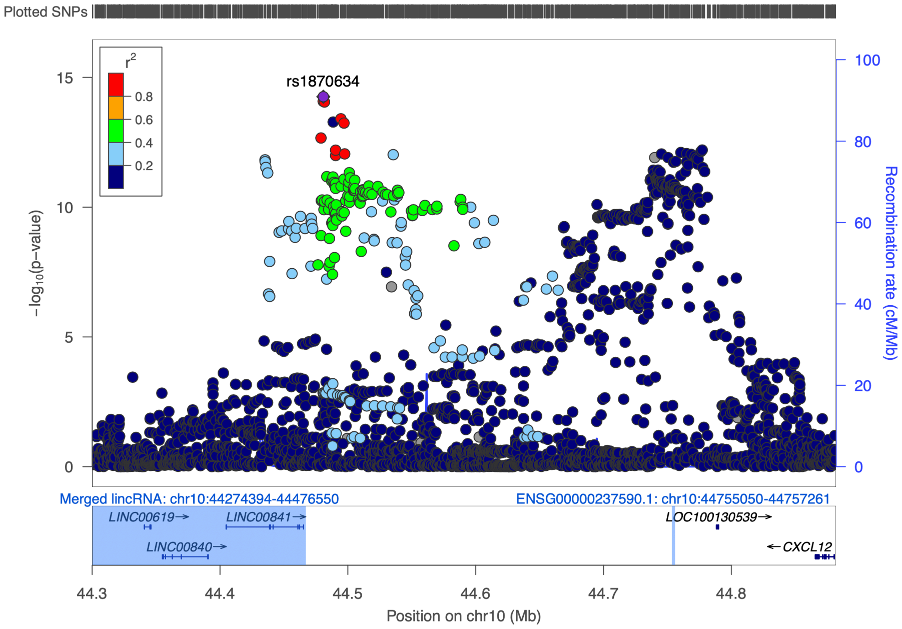
Supplement Figure III. Odds ratio (OR) of GWAS signals for conservation relative to non-conservation of lincRNAs based on multivariable models for two definitions of conservation



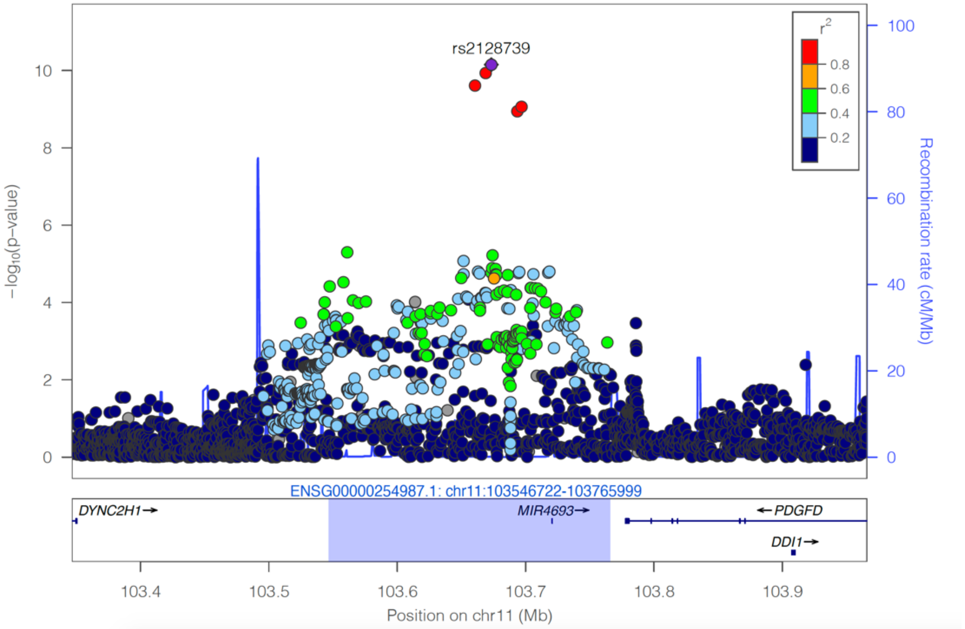
In this figure, “Syntenic” (light blue) indicates the OR of GWAS signals for conservation relative to non-conservation of lincRNAs based on the primary definition of conservation (syntenic vs. non-syntenic) and “Syntenic and expressed” (dark blue) indicates the OR of GWAS signals for conservation relative to non-conservation of lincRNAs based on the secondary definition of conservation (syntenic and expressed vs. non-syntenic or syntenic and not expressed).

Supplement Figure IV. Locus zoom plots of non-conserved (i and iii) and conserved (ii and iv) lincRNAs at loci with genome-wide significance for (A) coronary artery disease (CAD) and (B) waist-to-hip ratio adjusted for BMI (WHRadjBMI).

(A) Coronary artery disease (CAD)

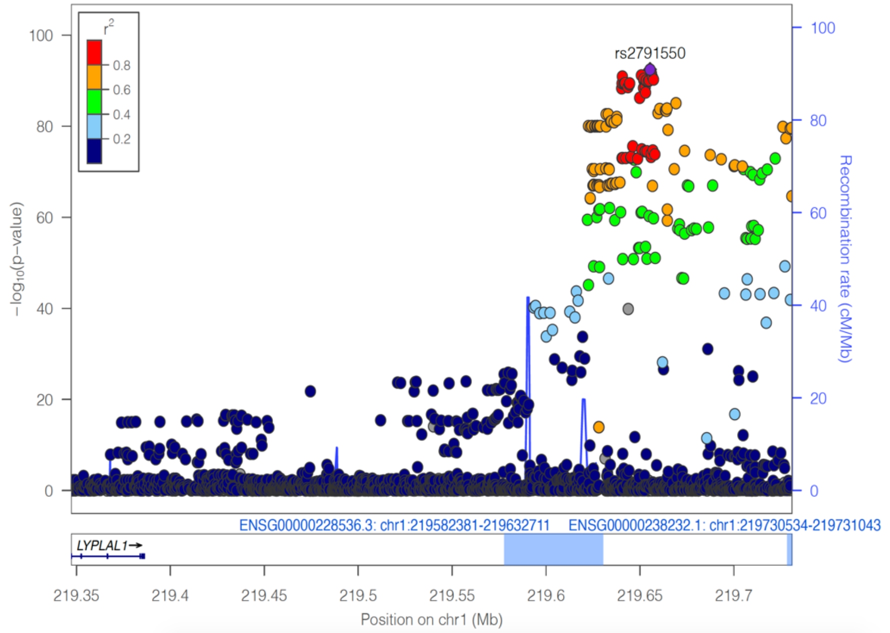


(i) Several non-conserved lincRNAs (merged lincRNA chr10:44274394-44476550 and ENSG00000237590) at the CXCL12 locus for CAD. Although CXCL12 has been implicated through functional studies as a potentially causal protein coding gene (PGC) at this locus, the non-conserved lincRNAs are candidate regulators of CXCL12 expression and CAD association at this locus.

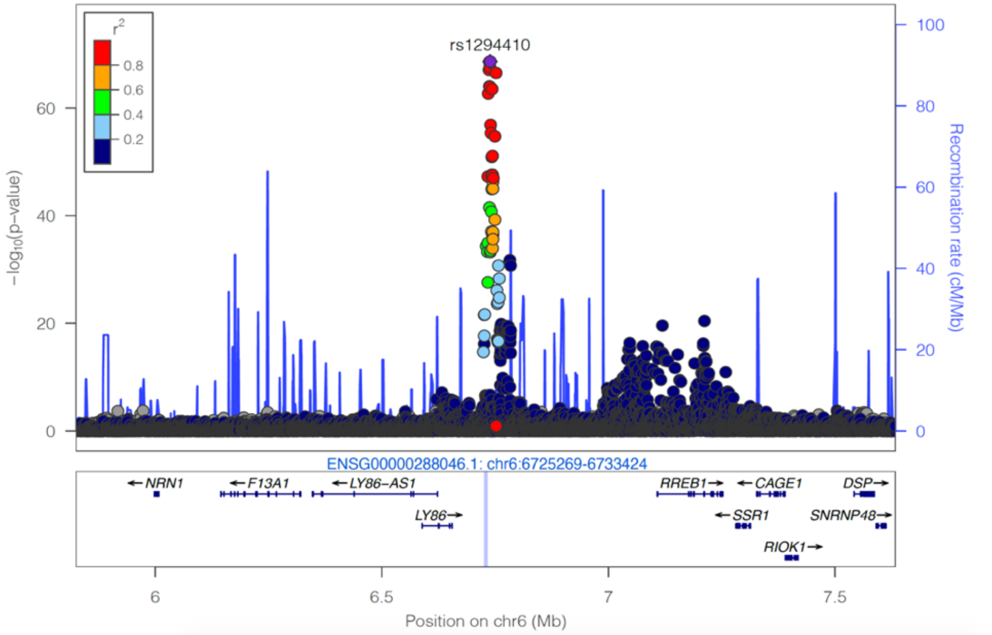


(ii) Conserved lincRNA (ENSG00000254987.1) at the PDGFD locus for CAD, along with 200kb upstream and downstream regions. Although PDGFD is a strong candidate PCG at this locus, it is downstream of the GWAS signal at the locus, while ENSG00000254987.1 overlaps the GWAS signal and is a candidate for CAD association, possibly via regulation of PDGFD expression.

(B) Waist to hip ratio adjusted for BMI (WHRadjBMI)

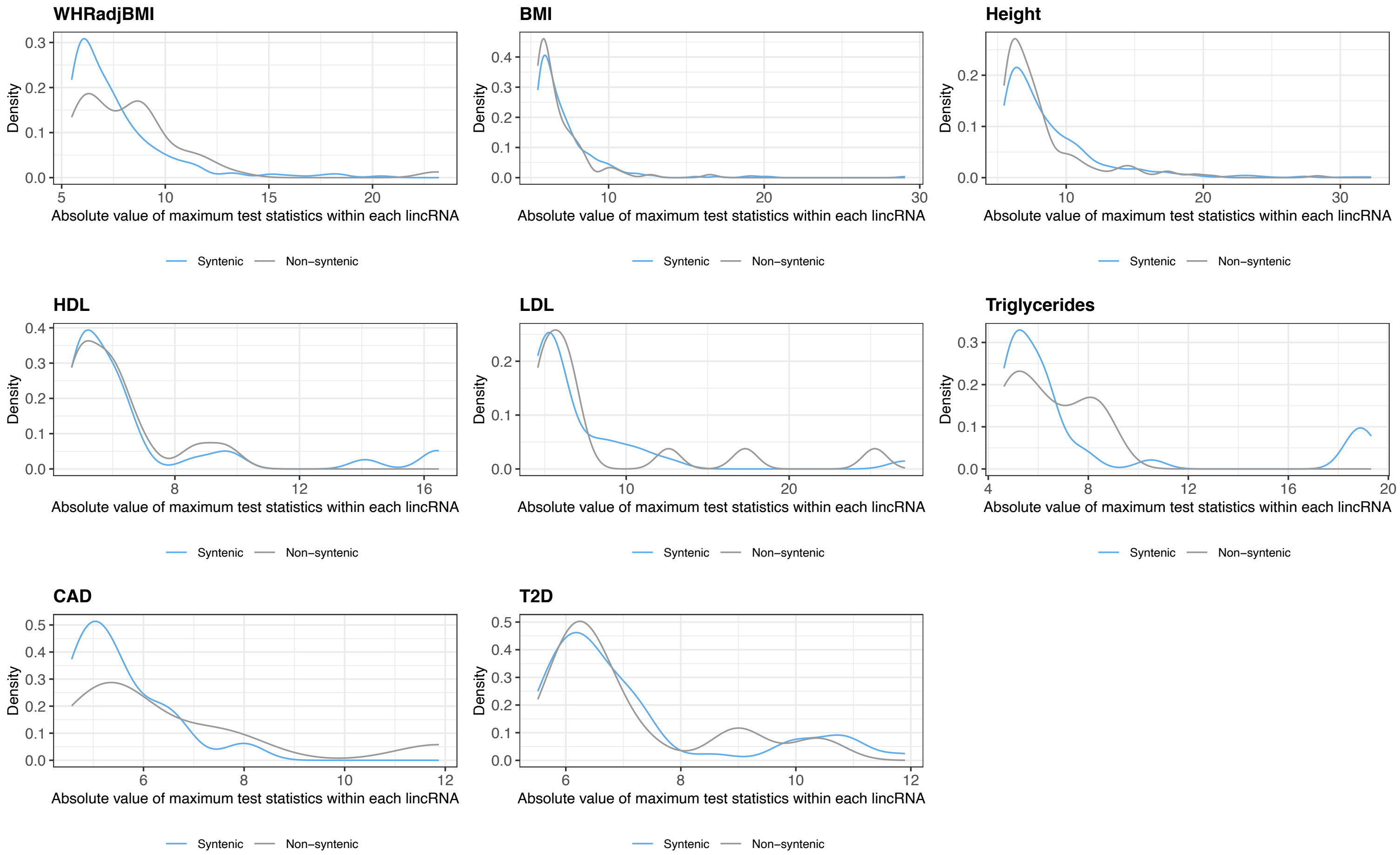


(i) Non-conserved lincRNAs (ENSG00000228536 and ENSG0000023832.1) at the LYPLAL1 locus for WHRadjBMI. Despite functional studies, regional PGCs, including LYPLAL1, have not been implicated as causal at this locus. The non-conserved lincRNAs are strong candidates for WHRadjBMIR association at this locus



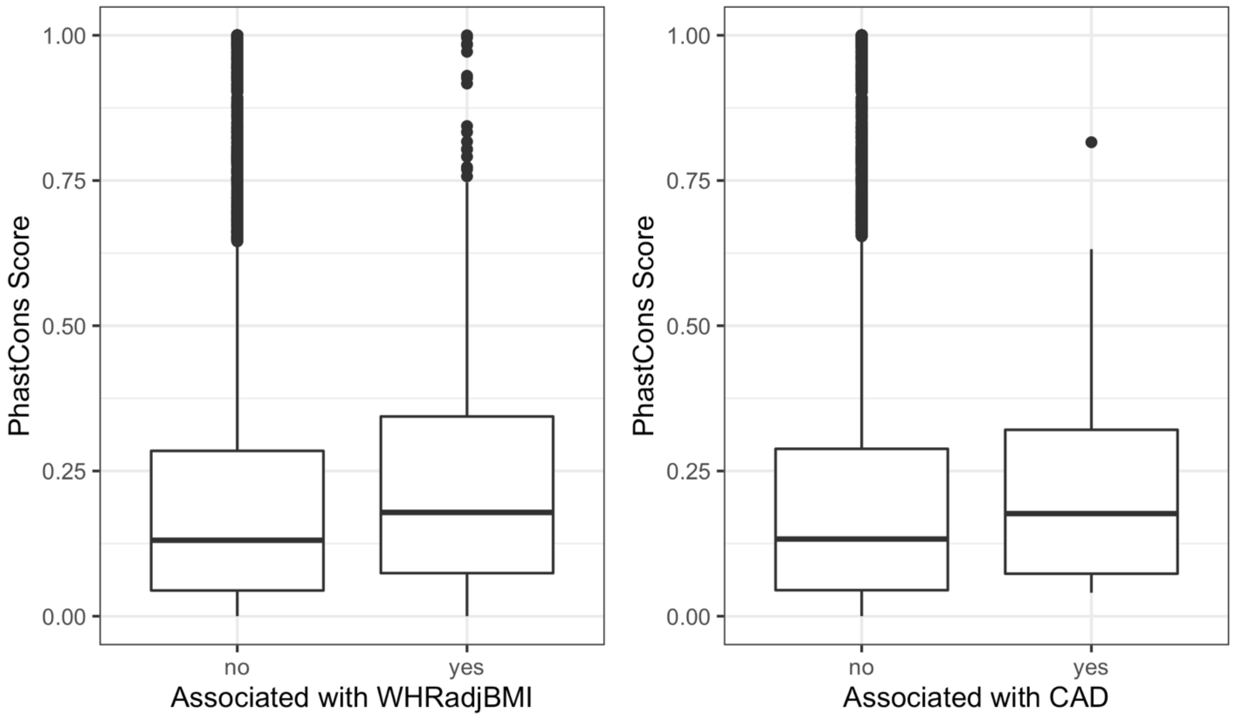
(ii) Conserved lincRNA (ENSG00000288046.1) at the LY86 locus for WHRadjBMI. There are multiple PCGs at this obesity locus but ENSG00000288046.1 is a strong candidate based on its proximity to and overlap with the GWAS signal.

Supplement Figure V. Truncated distributions of GWAS test statistics by the primary syntenic definition of conservation.



Density plots of the maximum absolute SNP-level z-score within signal lincRNAs. No apparent difference is observed in the density for conserved and non-conserved. This suggests that the magnitude of the signal in conserved lincRNAs is not greater than in non-conserved lincRNAs.

Supplement Figure VI. Distribution of phastCons scores for lincRNAs by GWAS signals for WHRadjBMI.



The median phastCons score is higher in lincRNAs associated with WHRadjBMI compared to lincRNAs not associated with WHRadBMI (Wilcoxon rank sum test p-value<0.001, left hand panel). There is no detectable difference in the median phastCons score for lincRNAs associated with CAD compared to lincRNAs not associated with CAD (Wilcoxon rank sum test p-value=0.310, right hand panel). Overall, the median phastCons scores for WHRadBMI and CAD associated lincRNAs are quite low (<0.2).

References

1. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13.
2. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols. 2009;4(1):44-57.