Arteriosclerosis, Thrombosis, and Vascular Biology

CLINICAL AND POPULATION STUDIES

Nonconserved Long Intergenic Noncoding RNAs Associate With Complex Cardiometabolic Disease Traits

Andrea S. Foulkes, Caitlin Selvaggi, Tingyi Cao, Marcella E. O'Reilly, Esther Cynn, Puyang Ma, Heidi Lumish, Chenyi Xue, Muredach P. Reilly

OBJECTIVE: Transcriptome profiling of human tissues has revealed thousands of long intergenic noncoding RNAs (lincRNAs) at loci identified through large-scale genome-wide studies for complex cardiometabolic traits. This raises the question of whether genetic variation at nonconserved lincRNAs has any systematic association with complex disease, and if so, how different this pattern is from conserved lincRNAs. We evaluated whether the associations between nonconserved lincRNAs and 8 complex cardiometabolic traits resemble or differ from the pattern of association for conserved lincRNAs.

APPROACH AND RESULTS: Our investigation of over 7000 lincRNA annotations from GENCODE Release 33–GRCh38.p13 for complex trait genetic associations leveraged several large, established meta-analyses genome-wide association study summary data resources, including GIANT (Genetic Investigation of Anthropometric Traits), UK Biobank, GLGC (Global Lipids Genetics Consortium), Cardiogram (Coronary Artery Disease Genome Wide Replication and Meta-Analysis), and DIAGRAM (Diabetes Genetics Replication and Meta-Analysis)/DIAMANTE (Diabetes Meta-Analysis of Trans-Ethnic Association Studies). These analyses revealed that (1) nonconserved lincRNAs associate with a range of cardiometabolic traits at a rate that is generally consistent with conserved lincRNAs; (2) these findings persist across different definitions of conservation; and (3) overall across all cardiometabolic traits, approximately one-third of genome-wide association study–associated lincRNAs are nonconserved, and this increases to about two-thirds using a more stringent definition of conservation.

CONCLUSIONS: These findings suggest that the traditional notion of conservation driving prioritization for functional and translational follow-up of complex cardiometabolic genomic discoveries may need to be revised in the context of the abundance of nonconserved long noncoding RNAs in the human genome and their apparent predilection to associate with complex cardiometabolic traits.

GRAPHIC ABSTRACT: A graphic abstract is available for this article.

Key Words: cardiovascular diseases ■ complex traits ■ Genome Wide Association Studies ■ lincRNAs ■ syntenic conservation

Most loci identified through large-scale genome-wide studies for complex cardiometabolic traits fall in intergenic regions, and many of these overlap genomic features that confer cell-specific regulatory functions. Indeed, transcriptome profiling of human tissues has revealed thousands of long intergenic noncoding RNAs (lincRNAs), representing the majority of all long noncoding RNAs (lncRNAs), transcribed in a cell- and tissue-specific manner at many of these loci raising the question as to whether these lincRNAs could be causal elements for cardiometabolic trait associations at these intergenic loci.[1] Convention in the field suggests that genetic elements that are conserved across many species are more likely to be functional and, if disrupted by mutations or common variation, contribute to rare diseases and complex traits, respectively. Evolutionary profiling shows that the majority of human lincRNAs mapped by RNA sequencing is not conserved outside of primate species, and it has been suggested that some proportion of these may not be true functional lncRNAs but rather byproducts of pervasive

Correspondence to: Andrea S. Foulkes, ScD, Massachusetts General Hospital, 50 Staniford St, Suite 560, Boston, MA 02114. Email afoulkes@mgh.harvard.edu
The Data Supplement is available with this article at https://www.ahajournals.org/doi/suppl/10.1161/ATVBAHA.120.315045.
For Sources of Funding and Disclosures, see page 510.

Arterioscler Thromb Vasc Biol is available at www.ahajournals.org/journal/atvb

Nonstandard Abbreviations and Acronyms

BMI	body mass index
CAD	coronary artery disease
CRISPR	clustered regularly interspaced short palindromic repeats
GWAS	genome-wide association study
HDL	high-density lipoprotein
LDL	low-density lipoprotein
LDL-C	low-density lipoprotein cholesterol
lincRNA	long intergenic noncoding RNA
lncRNA	long noncoding RNA
PCG	protein-coding gene
SNP	single-nucleotide polymorphism
TE	transposable element
WHRadjBMI	waist-to-hip ratio adjusted for body mass index

Highlights

- Of 1000s of long intergenic noncoding RNAs in the human genome, nonconserved long intergenic noncoding RNAs associate with cardiometabolic traits at a rate that is similar to that for conserved long intergenic noncoding RNAs.
- These findings are consistent across multiple cardiometabolic traits and persist using different definitions of conservation.
- For all cardiometabolic traits, more than one-third of genome-wide association study–associated long intergenic noncoding RNAs are nonconserved based on syntenic positional conservation, and this increases to as much as two-thirds using a more stringent definition of conservation.

transcription.[2–4] Recent work, however, is revealing many examples of nonconserved human lincRNAs that are functional and biologically important including a subset that may be the causal element at loci for human cardiometabolic and other diseases.[1,4–15]

These perspectives raise important questions as to whether genetic variation at nonconserved lincRNAs has any systematic association with complex cardiometabolic diseases, and if so, how different is this pattern from conserved lincRNAs. This is an important question in determining which human lincRNAs should be prioritized for functional and translational study. If nonconserved lincRNAs warrant systematic interrogation, this requires a shift in mindset and application of innovative in vivo humanized models to address the physiological roles and disease impact of nonconserved lincRNAs. More broadly, because human genomes contain mostly nonconserved lincRNAs, the traditional notion of conservation driving functional prioritization for mechanistic studies in cardiometabolic model systems may need revision in context of our expanding knowledge of diverse, nonconserved, functional regulatory features.

In the current work, we evaluated the likelihood that nonconserved lincRNAs have association with a complex cardiometabolic trait and whether this resembles or differs from the pattern of association for conserved lincRNAs. This included comprehensive consideration of summary data from multiple large meta-analyses genome-wide association study (GWAS) for 8 cardiometabolic disease–related traits: waist-to-hip ratio adjusted for body mass index (WHRadjBMI), body mass index (BMI),[16–20] height,[21] HDL (high-density lipoprotein) cholesterol, LDL (low-density lipoprotein) cholesterol (LDL-C), triglycerides,[22] coronary artery disease (CAD),[23] and type-2 diabetes.[24] For lincRNA interrogation, we utilized a well-defined and comprehensive set of over 7000 multi-exon lincRNAs that have been rigorously annotated (GENCODE Release 33–GRCh38.p13).[25] Conservation was defined using multiple distinct strategies, primarily based on the broad perspective of synteny, or positional genomic conservation,[2,3,26,27] with secondary incorporation of additional information on expression in mouse tissues[25] and an exploratory consideration of conservation based on base-pair sequence scoring.[28] Our analyses revealed that (1) nonconserved lincRNAs associate with cardiometabolic traits at a rate that is consistent with conserved lincRNAs; (2) these finding persist across different definitions of conservation; and (3) overall across all traits, approximately one-third of GWAS-associated lincRNAs are nonconserved, and this increases to about two-thirds using a more stringent definition of conservation.

MATERIALS AND METHODS

Anonymized data and materials are collected from existing publicly available repositories as indicated below.

Figure I in the Data Supplement provides a schematic illustration of how synteny is defined and summarizes our analytic pipeline. Key aspects of data preparation and analysis are described here. Additional Methods description including gene set enrichment analyses is provided in Material and Methods in the Data Supplement.

Data and Code Availability

All data used in the analyses contributing to this article are publicly available at the sites indicated. Code is available upon request to the corresponding author.

Conservation Determination

Conservation Based on Synteny

LincRNAs with expression in human tissues were identified using the Human Gencode v33 gene annotation build hg38 25 and were designated as intergenic if no protein-coding gene (PCG) start or stop locations were contained within the lincRNA start

and stop locations. For primary analysis, an lincRNA was conserved if it was syntenic and nonconserved if it was not syntenic. LincRNAs were defined as syntenic if (1) the nearest upstream and downstream neighboring PCGs in humans had one-to-one mouse homologs based on the Ensembl genome database release 47; and (2) the homologs were on the same chromosome with consistent relative orientation based on Mouse Gencode v24 25. Neighbors were defined as within 900 kb of the start and end position of the lincRNA as described previously.[2]

A subset (18.5%) of lincRNAs without 2 neighboring PCGs within this region were designated as absent neighbors and were not classified as either syntenic or nonsyntenic. Furthermore, for a small subset of lincRNAs (1.2%) that had upstream and downstream one-to-one mouse homologs but with inconsistent PCG orientation between human and mouse homologs, the lincRNA was designated as inconsistent orientation and not included in subsequent modeling. The relative orientation of mouse homologs for 2 lincRNAs could not be determined as they were not present in Mouse Gencode v24, and therefore, these lincRNAs were designated as having absent neighbors. This classification approach is summarized in the Figure I in the Data Supplement (top).

Conservation Based on Synteny and Expression

For secondary analysis, alternative definitions of lincRNA conservation were applied. An lincRNA was considered conserved if it was both syntenic at the genomic level in mouse and was also expressed in mouse tissues as defined by Mouse Gencode v24. An lincRNA was nonconserved if it was syntenic but not expressed in mouse tissues or not syntenic in mouse. Syntenic lincRNAs were defined as expressed if an lincRNA was present in mouse between the 2 identified PCG homologs based on Mouse Gencode v24. Syntenic lincRNAs with overlapping mouse homologs were unclassified under this definition of conservation.

Conservation Based on Sequence Scoring

Finally, as an exploratory analysis using a sequence-level approach to determining conservation, we calculated 7-way phastCons scores[28] for the 200-bp region of each lincRNA transcript's transcription start site. The average across the transcription start site region was calculated, and the maximum across all transcripts of an lincRNAs was used as an alternative measure of conservation.

Merging of lincRNAs

For focused interrogation of lincRNAs with GWAS summary data, lincRNA boundaries were extended by 5 kb, to include single-nucleotide polymorphisms (SNPs) in the canonical 5′ promoter and 3′ UTR (untranslated region) regulatory regions of lincRNAs. Resultant overlapping lincRNAs on the same strand were consolidated and treated as a single lincRNA. In the case that overlapping lincRNAs were on opposite strands, the lincRNA on the positive (+) strand was retained and the lincRNA on the negative (—) strand was removed. In merging lincRNAs, the following decision rules were applied: (1) if any of the merged lincRNAs were syntenic, the new merged lincRNA was classified as syntenic; (2) if none of the lincRNAs were syntenic but at least 1 was nonsyntenic, the new merged lincRNA was classified as nonsyntenic; (3) if none were syntenic or nonsyntenic but at least 1 had inconsistent orientation, the new merged lincRNA was classified as inconsistent orientation; and (4) in all remaining cases, new merged lincRNAs were classified as absent neighbors. For the secondary definition of conservation, a similar approach was applied where syntenic is replaced with syntenic and expressed and a final category is included based on overlapping homologs. Merged lincRNAs are removed from the exploratory analysis based on phastCons. The approach to merging lincRNAs is presented in Figure I in the Data Supplement (lower left).

GWAS Data Selection and Signal Determination

Cardiometabolic trait GWAS summary datasets with large participant numbers were selected to provide statistical power to evaluate the disease association of conserved versus nonconserved lincRNAs. Height was included because it is a defining complex genetic trait, has large GWAS sample sizes, and in recent years, height has been shown to share causal pathways with those for atherosclerotic cardiovascular disease.[29] SNPs were mapped from hg19 to hg38 using LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver). If the minimum SNP-level *P* within an lincRNA (±5 kb as outlined above) was less than the corresponding threshold (provided in Table 1), the lincRNA was classified as having a GWAS signal. For WHRadjBMI and BMI, the minimum *P* was determined across meta-analyses of men, women, and men and women combined. In all other cases, results were based on meta-analysis results for men and women combined. Guanine-cytosine (GC) content was calculated using sequence data from Human Gencode v33. For each lincRNA, exons were identified and merged if overlapping and GC content was defined as the proportion of Gs and Cs in the exon sequences. Transposable element (TE) coverage was defined as the proportion of the lincRNA exon sequences that overlap with TEs. For this calculation, the positions of TE types LINE (long interspersed nuclear element), SINE (short interspersed nuclear element), LTR (long terminal repeat), and DNA were identified using University of California Santa Cruz Genome Browser RepeatMasker.[30]

Statistical Analysis

The primary outcome was GWAS signal defined as an indicator that the minimum SNP-level *P* within the lincRNA (±5 kb as outlined above) was less than a predefined threshold (Table 1). Conservation was defined based on synteny for primary analysis and based on synteny and expression for secondary analysis. The proportions of conserved (·) and nonconserved (·) lincRNAs, respectively, with GWAS signal are reported. A noninferiority test given by $H_0: p_1 - p_2 < \delta$ versus $H_A: p_1 - p_2 < \delta$ is applied for each trait with $\delta = 0.01$. A corresponding $P<0.05$ was considered statistically meaningful and suggested that the proportion of nonconserved lincRNAs with GWAS signal was not significantly less than the proportion of conserved lincRNAs with GWAS signal in unadjusted analysis.

Additionally, multivariable logistic regression models were fitted separately for each trait and adjusted for the number of SNPs (natural log-transformed), GC content (natural log-transformed), and TE coverage. LincRNAs were treated as the unit of analysis, and data were limited to lincRNAs that were classified as conserved or not conserved. Wald tests of a difference in the probability of GWAS signal between conserved and nonconserved lincRNAs based on adjusted models are reported. Odds ratios and corresponding 95% CIs corresponding to

Table 1. Summary of Genome-Wide Association Study Data Resources

	No. of SNPs	Sample size	Coverage*	No. of LincRNAs mapped to SNPs	Signal threshold†	Source
WHRadjBMI	27 364 379	694 649	183 (123.5–388)	7011	5×10^{-8}	GIANT/UKBb[18]
BMI	27 369 701	806 834	183 (124–388.5)	7011	5×10^{-8}	GIANT/UKBb[18]
Height	2 332 944	≈700 000	16 (8–37)	6611	5×10^{-8}	GIANT/UKBb[21]
HDL	2 445 954	188 577	17 (8–39)	6704	5×10^{-6}	GLGC[22]
LDL	2 437 751	188 577	17 (8–39)	6698	5×10^{-6}	GLGC[22]
TGs	2 439 264	188 577	17 (8–39)	6698	5×10^{-6}	GLGC[22]
CAD	9 455 778	184 305	67 (40–142)	6859	5×10^{-6}	Cardiogram[23]
T2D	21 635 866	898 130	146 (98–309)	6977	5×10^{-8}	DIAGRAM[24]

*Median number of SNPs per lincRNA and interquartile range (25th to 75th percentile).

†Signal threshold was set to 5×10^{-8} for analysis of GIANT/UKBb and DIAGRAM data to correct for multiple comparisons. A less stringent but still suggestive threshold of 5×10^{-6} was used for the analysis of GLGC and cardiogram data as the sample size, and, therefore, the power for detecting association is lower in these settings. BMI indicates body mass index; CAD, coronary artery disease; Cardiogram, Coronary Artery Disease Genome Wide Replication and Meta-Analysis; DIAGRAM, Diabetes Genetics Replication and Meta-Analysis; GIANT, Genetic Investigation of Anthropometric Traits; GLGC, Global Lipids Genetics Consortium; HDL, high-density lipoprotein; LDL, low-density lipoprotein; LincRNA, long intergenic noncoding RNA; SNP, single-nucleotide polymorphism; T2D, type 2 diabetes; TG, triglyceride; UKBb, UK Biobank; and WHRadjBMI, waist-to-hip ratio adjusted for body mass index.

the odds of GWAS signal for conserved lincRNAs compared with the odds of GWAS signal for nonconserved lincRNAs are also provided. The estimated probabilities of GWAS signal for conserved and nonconserved lincRNAs and the corresponding prediction interval were determined based on the multivariable fitted logistic model. A summary of the statistical analysis approach is provided in Figure I in the Data Supplement (right).

RESULTS

Descriptive Characteristics of lincRNAs

Publicly available GWAS summary data used in the analysis are summarized in Table 1 and included WHRadjBMI and BMI,[18] height,[21] HDL cholesterol, LDL-C, and triglycerides,[22] CAD,[23] and type-2 diabetes.[24] Table 2 illustrates the distributions of transcript length, GC content, exon count, and TE coverage for lincRNAs that are classified as conserved or nonconserved (defined based on synteny as described in Methods). Summary data are reported as medians and interquartile ranges as these measures are robust to skewness in the data. Information on lincRNAs unclassified due to the absence of a PCG upstream or downstream (or both) or with inconsistent PCG relative orientation is provided in Table I in the Data Supplement.

Conserved lincRNAs tend to be longer than nonconserved lincRNAs, and this difference is more pronounced when expression in mouse is considered (median length: syntenic lincRNAs, 15960 bps; nonsyntenic lincRNAs, 15851 bps; syntenic and expressed lincRNAs, 19120 bps; syntenic and not expressed or nonsyntenic lincRNAs, 14721 bps). Moreover, unclassified lincRNAs tend to be significantly longer with a lower GC content and higher TE coverage (Table I in the Data Supplement). The number of SNPs per lincRNA (based on WHRadjBMI data) tracks with the length of the lincRNA, so that the distribution of the number of SNPs divided by lincRNA length is approximately the same in all categories. Overall, these findings support the use of multivariable adjusted analyses including these variables as potential confounders in characterizing the relationship between lincRNA conservation and GWAS signal.

Table 2. Characteristics of Conserved and Nonconserved lincRNAs

	Conservation based on synteny		Conservation based on synteny and expression		
Characteristic*	Conserved (n=4243)	Nonconserved (n=1445)	Conserved (n=2262)	Nonconserved (n=3398)	Total (n=7089)
Length	15960 (11918–32571)	15851 (11984–29670)	19120 (12510–44262)	14721 (11702–26402)	17130 (12093–36922)
GC content	0.458 (0.416–0.506)	0.460 (0.415–0.504)	0.452 (0.413–0.498)	0.463 (0.418–0.509)	0.450 (0.407–0.499)
Exon count	3 (2–4)	3 (2–4)	3 (2–5)	2 (2–4)	3 (2–5)
TE coverage	0.346 (0.154–0.543)	0.329 (0.130–0.541)	0.334 (0.149–0.518)	0.347 (0.147–0.558)	0.346 (0.155–0.545)
No. of SNPs†	178 (123–354)	161 (108–293)	214 (136–484)	157 (111–272)	183 (123.5–388)
No. of SNPs/length†	0.010 (0.009–0.012)	0.010 (0.008–0.012)	0.011 (0.009–0.012)	0.010 (0.009–0.012)	0.010 (0.009–0.012)

GC indicates guanine-cytosine; GWAS, genome-wide association study; lincRNA, long intergenic noncoding RNA; SNP, single-nucleotide polymorphism; TE, transposable element; and WHRadjBMI, waist-to-hip ratio adjusted for body mass index.

*Median and interquartile range (25th to 75th percentile) across lincRNAs within corresponding category.

†Summary results for the number of SNPs per lincRNA and number of SNPs divided by lincRNA length are based on a subset of n=7011 lincRNAs and GWAS SNPs for WHRadjBMI (Table 1).

Analysis Using Primary Definition of Conservation Based on Synteny

The counts and percentages of lincRNAs by conservation and GWAS signal are provided in Table 3. In this unadjusted analysis based on the primary definition of conservation, the estimated proportion of lincRNAs with GWAS signal for nonconserved lincRNAs is less than the corresponding proportion for conserved lincRNAs for BMI (6.2% versus 6.9%; noninferiority $P>0.05$) and height (16.8% versus 18.8%; noninferiority $P>0.05$) while this estimated proportion is greater in nonconserved compared with conserved lincRNAs for WHRadjBMI (5.7% versus 5.0%; noninferiority $P<0.01$), HDL cholesterol (1.0% versus 0.7%; noninferiority $P<0.001$), LDL-C (1.2% versus 0.6%; noninferiority $P<0.001$), triglycerides (1.1% versus 0.7%; noninferiority $P<0.001$), CAD (0.6% versus 0.4%; noninferiority $P<0.001$), and type-2 diabetes (1.6% versus 1.1%; noninferiority $P<0.001$).

Overall, these findings suggest that it is as likely for nonconserved lincRNAs as for conserved lincRNAs to include a GWAS-associated SNP. In addition, a

Table 3. Genome-Wide Association Study Signal Counts by Trait and Conservation (Unadjusted Analysis)

		No signal	Signal (column percentage)	Total	Signal, %	Test of non-inferiority*
Conservation defined based on synteny						
WHRadjBMI (n=5635)	Nonconserved	1315	80 (27.6%)	1395	5.7	0.00796
	Conserved	4030	210	4240	5.0	
BMI (n=5635)	Nonconserved	1308	87 (23.0%)	1395	6.2	0.345
	Conserved	3949	291	4240	6.9	
Height (n=5319)	Nonconserved	968	195 (20.0%)	1163	16.8	0.212
	Conserved	3375	781	4156	18.8	
HDL (n=5395)	Nonconserved	1208	12 (29.3%)	1220	1.0	<0.001
	Conserved	4146	29	4175	0.7	
LDL (n=5389)	Nonconserved	1203	14 (36.8%)	1217	1.2	<0.001
	Conserved	4148	24	4172	0.6	
TGs (n=5389)	Nonconserved	1204	13 (29.5%)	1217	1.1	<0.001
	Conserved	4141	31	4172	0.7	
CAD (n=5534)	Nonconserved	1301	8 (33.3%)	1309	0.6	<0.001
	Conserved	4209	16	4225	0.4	
T2D (n=5616)	Nonconserved	1354	22 (31.4%)	1376	1.6	<0.001
	Conserved	4192	48	4240	1.1	
Conservation defined based on synteny and expression						
WHRadjBMI (n=5607)	Nonconserved	3173	173 (59.9%)	3346	5.2	0.0336
	Conserved	2145	116	2261	5.1	
BMI (n=5607)	Nonconserved	3115	231 (61.3%)	3346	6.9	0.0196
	Conserved	2115	146	2261	6.5	
Height (n=5292)	Nonconserved	2481	579 (59.6%)	3060	18.9	0.0160
	Conserved	1840	392	2232	17.6	
HDL (n=5368)	Nonconserved	3100	28 (68.3%)	3128	0.9	<0.001
	Conserved	2227	13	2240	0.6	
LDL (n=5362)	Nonconserved	3097	26 (68.4%)	3123	0.8	<0.001
	Conserved	2227	12	2239	0.5	
TGs (n=5362)	Nonconserved	3094	29 (65.9%)	3123	0.9	<0.001
	Conserved	2224	15	2239	0.7	
CAD (n=5506)	Nonconserved	3233	17 (70.8%)	3250	0.5	<0.001
	Conserved	2249	7	2256	0.3	
T2D (n=5588)	Nonconserved	3280	47 (68.1%)	3327	1.4	<0.001
	Conserved	2239	22	2261	1.0	

BMI indicates body mass index; CAD, coronary artery disease; HDL, high-density lipoprotein; LDL, low-density lipoprotein; T2D, type 2 diabetes; TG, triglyceride; and WHRadjBMI, waist-to-hip ratio adjusted for body mass index.

*Test of noninferiority is based on $\delta=0.01$.

substantial number of GWAS-associated lincRNAs are not conserved, as indicated by the column percentages in Table II in the Data Supplement. For example, 80 of 290 (27.6%) lincRNAs with a GWAS signal for WHRadjBMI are nonconserved. The percentage of GWAS-associated lincRNAs that are not conserved ranges from 20% (for height) to 36.8% (for LDL-C).

Multivariable Models

The results of multivariable modeling (Table 4; Figure; Figures II and III in the Data Supplement) are consistent with findings of unadjusted analyses with the exception that the predicted probability of GWAS signal for BMI is now slightly higher for nonconserved compared with conserved lincRNAs. The corresponding adjusted estimated odds ratio of conserved, relative to nonconserved, lincRNA association with traits is <1 for all traits except height ($P<0.05$ for LDL-C; $P>0.05$ for all other traits) and ranges from 0.451 (95% CI, 0.231–0.878) for LDL to 1.126 (95% CI, 0.947–1.338) for height.

Illustrative Examples

As illustrative examples, Figure IV in the Data Supplement presents locus plots for several examples of genetic loci containing nonconserved and conserved lincRNAs that are associated with CAD and WHRadjBMI—2 well-studied and clinically important cardiometabolic traits.

Secondary Analysis Using Alternative Definitions of Conservation

Using the secondary definition of conservation that requires lincRNA expression in mouse and human, as well as synteny, the predicted probability of GWAS signal is higher in nonconserved lincRNAs compared with conserved lincRNAs for all traits ($P<0.05$ for BMI, height, and LDL-C; $P>0.05$ for all other traits; Table 4; Figure III in the Data Supplement). Notably, for both definitions of conservation, the point estimate for the probability of GWAS signal is consistently greater in nonconserved lincRNAs compared with conserved lincRNAs. Although this difference is not statistically significant for most traits considered individually, the overall trend suggests

Table 4. Multivariable Adjusted Model Estimates for Effect of Conservation on GWAS Signal by Trait

	Estimate for syntenic*	SE	*Z* value	Pr(>|z|)	OR (95% CI)
Conservation defined based on synteny					
WHR-adjBMI	−0.220	0.136	−1.613	0.107	0.803 (0.614–1.048)
BMI	−0.005	0.391	−0.040	0.968	0.995 (0.775–1.277)
Height	0.118	0.088	1.343	0.179	1.126 (0.947–1.338)
HDL	−0.412	0.346	−1.189	0.235	0.663 (0.336–1.306)
LDL	−0.796	0.340	−2.341	0.019	0.451 (0.231–0.878)
TG	−0.442	0.334	−1.325	0.185	0.643 (0.334–1.236)
CAD	−0.546	0.436	−1.253	0.210	0.579 (0.247–1.361)
T2D	−0.415	0.262	−1.589	0.112	0.660 (0.395–1.102)
Conservation defined based on synteny and expression					
	Estimate for syntenic and expressed*				
WHR-adjBMI	−0.085	0.127	−0.667	0.505	0.919 (0.716–1.179)
BMI	−0.276	0.114	−2.419	0.016	0.759 (0.607–0.949)
Height	−0.155	0.075	−2.077	0.038	0.856 (0.740–0.991)
HDL	−0.634	0.348	−1.820	0.069	0.530 (0.268–1.050)
LDL	−0.745	0.364	−2.047	0.041	0.475 (0.233–0.969)
TG	−0.551	0.331	−1.664	0.096	0.576 (0.301–1.103)
CAD	−0.761	0.462	−1.648	0.099	0.467 (0.189–1.155)
T2D	−0.488	0.267	−1.825	0.068	0.614 (0.363–1.037)

BMI indicates body mass index; CAD, coronary artery disease; GWAS, genome-wide association study; HDL, high-density lipoprotein; LDL, low-density lipoprotein; lincRNA, long intergenic noncoding RNA; OR, odds ratio; SNP, single-nucleotide polymorphism; T2D, type 2 diabetes; TE, transposable element; TG, triglyceride; and WHRadjBMI, waist-to-hip ratio adjusted for body mass index.

*Separate multivariable models are fitted for each trait. Models are adjusted for number of SNPs (natural log-transformed), guanine-cytosine (GC) content (natural log-transformed), and TE coverage. In the model for the WHRadjBMI signal with conservation defined based on synteny, the OR corresponding to a 1-unit change in natural log GC content is 3.20 ([95% CI, 1.36–7.48] $P=0.007$) and the OR for 1 unit change in TE coverage is 0.837 ([95% CI, 0.522–1.35] $P=0.459$). This suggests that GC content is significantly associated with the probability of a GWAS signal for WHRadjBMI. Adjustment for these additional covariates supports the unadjusted finding that the likelihood for a nonconserved lincRNA to include a GWAS signal SNP is similar to that of a conserved lincRNA.

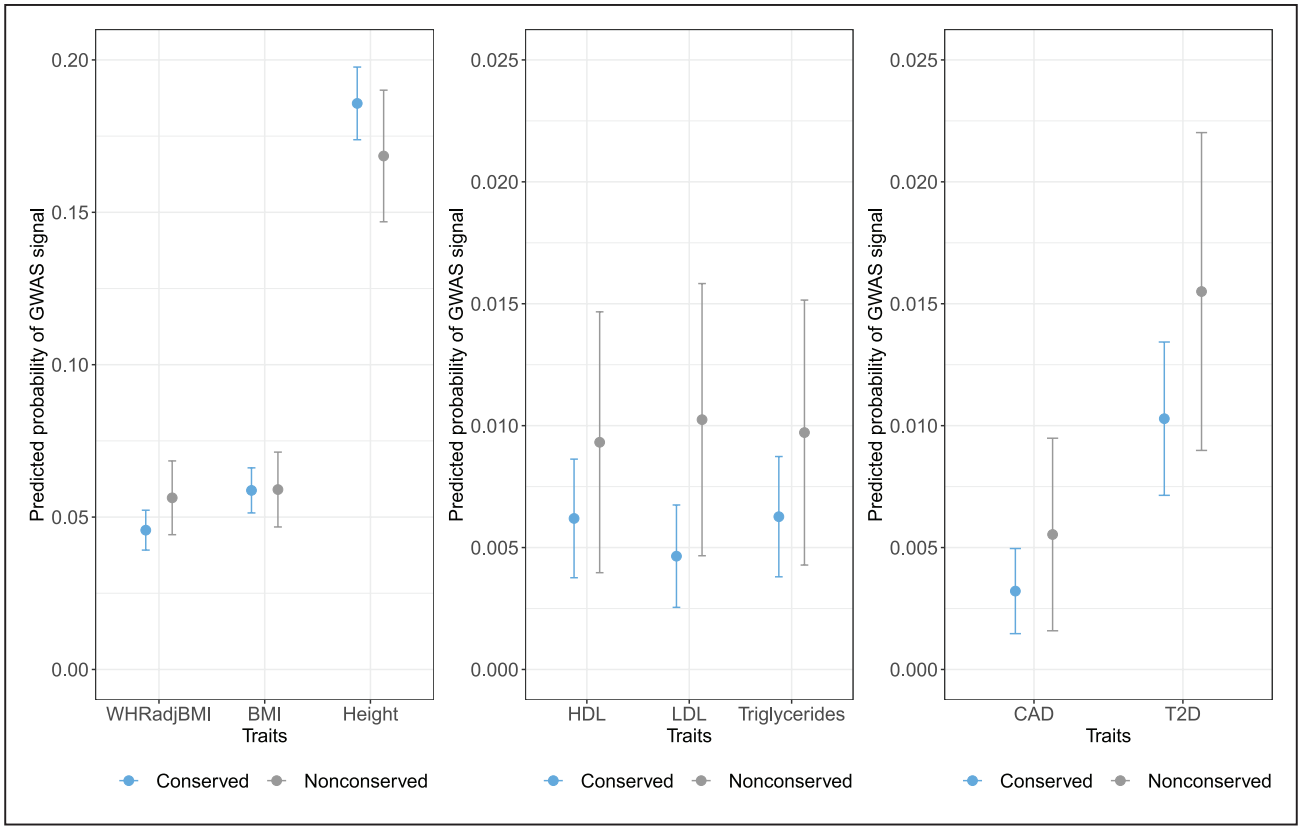

Figure. Predicted probabilities of genome-wide association study (GWAS) signal for conserved and nonconserved long intergenic noncoding RNAs (lincRNAs).

Predicted probabilities and corresponding 95% prediction intervals are calculated based on multivariable models using average of observed median values for guanine-cytosine content and transposable element coverage and observed trait-specific median number of single-nucleotide polymorphisms. The predicted probability of GWAS signal is greater for nonconserved lincRNAs than conserved lincRNAs for all traits considered except height based on the primary definition of conservation. The results based on the secondary definition of conservation are consistent though in this case, the predicted probably of GWAS signal is greater for nonconserved lincRNAs than conserved lincRNAs for all traits including height (results not shown). The consistently overlapping CIs suggest that the likelihood of GWAS association for conserved and nonconserved lincRNAs is comparable, and, therefore, the traditional metrics of conservation for prioritizing long noncoding RNAs for functional studies needs to be reconsidered. CAD indicates coronary artery disease; HDL, high-density lipoprotein; LDL, low-density lipoprotein; T2D, type-2 diabetes; and WHRadjBMI, waist-to-hip ratio adjusted for body mass index.

that the notion that GWAS signal would be lower in nonconserved regions needs to be reconsidered. Similar to the first definition of conservation, a substantial number of GWAS-associated lincRNAs are not conserved based on the secondary definition (Table II in the Data Supplement). In this case, 173 of 289 (59.9%) GWAS-associated lincRNAs for WHRadjBMI are nonconserved. This percentage of GWAS-associated lincRNAs that are not conserved, based on the secondary definition, ranges from 59.6% (for height) to 70.8% (for CAD).

Additional Analyses

To compare the strength of lincRNA GWAS signals, we plotted the density of the maximum within lincRNA SNP-level *Z* score among trait-associated lincRNAs for conserved and nonconserved lincRNAs using our primary syntenic definition of conservation (Figure V in the Data Supplement). No apparent trend is observed to suggest that the magnitude of the association signal in conserved lincRNAs is greater than nonconserved lincRNAs.

To probe features of lincRNAs that were unclassified in our primary syntenic definition of conservation (ie, the 18.5% lincRNAs that lack 2 neighboring PCGs within 900 kb of their start and end positions), counts and associated models comparing the set of unclassified lincRNAs to lincRNAs that are classified as either conserved or nonconserved are provided in Tables III and IV in the Data Supplement. These results generally suggest a lower probability of GWAS signal in more isolated genomic regions within which the majority of unclassified lincRNAs is found.

In exploratory analysis of sequence-level conservation, the distribution of lincRNA level phastCons scores by GWAS association for WHRadjBMI and CAD are provided in Figure VI in the Data Supplement. For lincRNAs associated with compared with lincRNAs not associated with these traits, the median phastCons score is higher in

lincRNAs associated with WHRadjBMI (Wilxocon rank-sum test, $P<0.001$, left) but not lincRNAs associated with CAD (Wilxocon rank-sum test, $P=0.310$, right). Although there is a statistically significant difference in the median phastCons score for WHRadBMI, the distribution of phastCons for WHRadjBMI-associated lincRNAs ranges from 0 to 1 with a large proportion of relatively low scores and a low average phastCons score for WHRadjBMI, as well as for CAD-associated lincRNAs.

To explore lincRNA regulatory and functional features, we examined whether neighboring PCGs of conserved and nonconserved disease-associated lincRNAs were enriched in different pathways that might hint at differences in their regulatory functions in cardiometabolic traits. Using WHRadjBMI as an example, we performed pathway-based analysis using the Database for Annotation, Visualization and Integrated Discovery (https://david.ncifcrf.gov/)[31,32] based on neighboring PCGs of trait-associated conserved and nonconserved lincRNAs. Each interrogation of the Database for Annotation, Visualization and Integrated Discovery categories showed similar findings, so we present the results from UniProt Keyword (UP_Keyword) annotations in Table V in the Data Supplement. For WHRadjBMI-associated lincRNAs, biological processes were different for PGCs at conserved versus those at nonconserved lincRNAs—PCG neighbors of conserved lincRNA are significantly enriched in transcriptional regulation and DNA binding, whereas PCG neighbors of nonconserved lincRNA enrich for major histocompatibility complex I, immunity, and cell division.

DISCUSSION

A large portion of human lncRNAs lack conservation; yet, emerging evidence suggests nonconserved lncRNAs are functional.[1,4–15,26,33,34] Motivated by this, we evaluated the likelihood that nonconserved lincRNA loci have genetic association with complex human cardiometabolic traits and compared this to the pattern of association for conserved lincRNAs. Focusing on 8 established cardiometabolic disease–related traits,[35,36] we found that nonconserved lincRNAs have a similar likelihood of associating with cardiometabolic traits as conserved lincRNAs and that this association was broadly consistent across different definitions of conservation and different cardiometabolic traits. Moreover, approximately one-third of trait-associated lincRNA loci were nonconserved based on a syntenic definition of conservation and closer to two-thirds were not conserved based on a more rigorous definition that included both synteny and expression in mouse. These findings suggest that the traditional notion of conservation driving prioritization for functional and translational follow-up of human cardiometabolic genomic discoveries may need to be revised in the context of the abundance of nonconserved lincRNAs in the human genome and their apparent predilection to associate with complex disease traits.

Species conservation, at DNA and protein sequence levels, has been considered an important feature, and often used for primary triage, when determining whether a PCG is likely to be functional. This perspective is reinforced by decades of using model organisms, particularly mouse genetic models, relative to human or primate studies, to study in vivo function. However, a primary focus on conservation and use of mouse models may be to deprioritize important genetic signals for human diseases when considering genomic and regulatory features, including alternative splicing, tissue-specific enhancers, and lincRNAs, that are prominent features of primate evolution.[37] Although the protein-coding genome is largely conserved between primates and nonprimates, many cell-specific regulatory features are not conserved outside primates. This should not be altogether surprising because the specialized cell and organ functions that have emerged with primate evolution cannot be explained by changes in numbers of PCGs. This lack of conservation is particularly marked for lincRNAs, and our work[26,27] and that of others[2,3] suggests that the majority of human lincRNAs is not conserved in mice.

An alternative measure of conservation that is applied to PCGs is base-pair sequence homology.[4,38] However, human lincRNAs that are syntenic, expressed in mouse tissues, and functionally conserved often have limited nucleotide sequence homology across species.[2–4] For this reason, we focus in this work on genomic synteny between human and mouse as a primary measure of conservation. In our exploratory analysis of sequence conservation, while the central tendency of phastCons scores is higher in WHRadjBMI-associated lincRNAs compared with nonassociated lincRNAs, the low average phastCons score for WHRadjBMI- and CAD-associated lincRNAs, relative to PCGs, confirms a low sequence-level conservation for trait-associated lincRNAs. This suggests poor utility of sequence-level conservation scores in discriminating disease-associated from non–disease-associated lincRNAs.

While it has been proposed that many nonconserved lncRNA molecules that are identified through RNA sequencing technologies may be nonfunctional, several lines of evidence suggest that this is not the case. Genomic markers of function including tissue-enrichment, binding of tissue-specific transcription factors at lncRNA enhancers and promoters, and regulation in response to physiological stressors do not differ significantly between conserved and nonconserved myeloid and other tissue lincRNAs.[26,27,39,40] Several groups have also published genomic criteria, not dependent simply on conservation, and experimental methods, including CRISPR (clustered regularly interspaced short palindromic repeats) screens, to predict lncRNA functionality and prioritize candidates.[1,4,5,9] Multiple examples have

emerged of lincRNAs that overlap loci for human cardiometabolic traits,[1,12,15] including *ANRIL*, *H19*, *MALAT1*, *MEXIS*, *LOC157273*, and *LASER*.[6–8,10,11,13,14] Of these, there are several examples of conserved (syntenic) lincRNAs including *MALAT1* and *LOC157273* (*RP11-10A14.4*). There are also examples of functionally characterized nonconserved lincRNAs at loci for cardiometabolic disease traits despite limited functional studies including H19, which also has been shown to have higher plasma levels of H19 in patients with CAD.[34]

In a recent prepublication, the GTEx (Genotype-Tissue Expression) consortium performed colocalization analysis connecting genetic variation, gene expression, and traits for a set of 690 human lncRNAs by integrating results from GWAS for 48 traits and expression quantitative trait loci for 48 tissues in the latest GTEx v8 data.[1] Of 4694 significant expression quantitative trait loci GWAS SNP colocalization events for these lncRNAs and traits, a striking 80% lacked any colocalization with PCGs.[1] Although the GTEx work did not focus on measures of lncRNA conservation, our current findings suggest that a large proportion of lncRNAs that colocalize at loci for complex cardiometabolic traits lacks conservation in mice. Further, many primate-specific lincRNAs, not found in rodents or other model organisms, have emerged as important regulators in cellular processes, such as pluripotency and differentiation, and as noted above, several have been implicated in human cardiometabolic disorders.[5,26,41–44] These data and our exploratory finding of differences in gene-pathway enrichment for neighboring PCGs suggest there may be utility in considering regulatory and functional features, as well as disease association, rather than an initial triage using conservation, to identify and prioritize human lincRNAs for translational study.

A reluctance to study nonconserved lncRNAs also may hamper the development of rigorous and reproducible model systems to address pathophysiological functions of nonconserved lncRNAs and other genomic elements. Recent advances in tissue engineering have established stem cell–based organoids as near-physiological systems to study human physiology and diseases.[45,46] Modulation of PCGs and microRNAs by RNAi (RNA interference) or transgene have been used in nonhuman primates in translational or preclinical studies. However, nonhuman primates are scarce and costly, limiting feasibility. Much work on functional models is needed including transgene approaches that can express primate-specific lincRNAs in nonprimate animal models—indeed, a few studies show that protein or RNA partners of such lincRNAs are conserved and can interact with primate-specific lincRNAs in nonprimate models.[41] Bacterial artificial chromosome transgene mouse models can include the gene body and large fragments of genomic regulatory DNA of nonconserved lincRNA loci to drive human lincRNA expression in mouse models in vivo.[47] An additional in vivo approach is to engraft human cells expressing primate-specific lincRNAs in rodent models with immune deficiency as has been used to study the roles of human lincRNA in tumor development and metastasis.[48,49]

In our analyses, a substantial subset of lincRNAs (18.5%) were characterized as unclassified in terms of synteny because they lacked PCG within the published range of 900 kb[2] that we applied to examine PCGs upstream or downstream of a given lincRNA. These unclassified lincRNAs tend to be longer with a lower GC content and higher TE coverage relative to classified lincRNAs (Table I in the Data Supplement). Using an established minimum-range cut point for gene deserts of absence of a PCG within 250 kb upstream and 250 kb downstream,[50–52] 55.1% of unclassified compared with 7.3% of classified lincRNAs reside within gene deserts. Gene deserts, and lincRNAs within such regions, are enriched in ancient duplications, have lower GC content and lower conservation than other parts of the human genome, and may have specific long-distance *cis*- and *trans*-regulatory functions related to their unique evolutionary and genomic characteristics.[50–52] Although unrelated to our primary focus on the role of lincRNA conservation in human complex diseases, further study of these unique unclassified lincRNAs in gene deserts is of interest to the field. Indeed, there are well-recognized loci in gene desert that associate with complex traits at GWAS including the 9p21 locus with CAD and type-2 diabetes[53] and the 8q24 locus with several cancers .[54] Our analyses, however, suggest a lower probability of GWAS signal for unclassified lincRNAs that lie in more isolated genomic regions and gene deserts compared with classified lincRNAs (Tables III and IV in the Data Supplement).

Our study has several limitations. For example, there are no established standards in the field regarding the definition of lincRNA conservation, and, therefore, we chose somewhat arbitrary, although previously published,[2,3,26,27] definitions of synteny. For example, we excluded certain lincRNAs that lacked PCGs within 900 kb of lincRNAs. We also merged overlapping lincRNAs, and this may not accurately reflect the precise lincRNA and isoform expression in individual tissues or across tissues. Although GENCODE as a resource for lincRNAs is widely used and well cross-validated, it may lack sensitivity to many lncRNAs as expression of some functional lincRNAs can be highly context specific and found at low levels and therefore missed in the GENCODE resource. Indeed, our group[26,27] and others[55] have published such findings in several prior articles. Although our trait selection is comprehensive, we did not interrogate an all-encompassing set of cardiometabolic traits. Rather, we focused primarily on traits with adequately powered GWAS datasets that provided sufficient numbers of trait-associated SNPs in both conserved and nonconserved lincRNAs. In addition, our use of large SNP-based GWAS

datasets rather than whole genome data did not permit interrogation or rare functional variation and lincRNA exonic regions and did not provide the level of coverage required for a fine-mapping subset analysis focused on SNPs within exons and introns. As larger whole genome datasets emerge, there will be opportunities to focus on rare functional variations in lincRNAs, as well as analysis that can weight for enriched signals in 5′, 3′, exonic, and intronic SNPs and regions of lincRNAs.

In conclusion, we found that nonconserved lincRNAs have a nontrivial and consistent likelihood of association with a broad array of complex cardiometabolic traits. Indeed, we found that nonconserved lincRNAs associate with cardiometabolic traits at a rate that is consistent with conserved lincRNAs, that these findings are robust across different definitions of conservation, and strikingly that across all traits as much as two-thirds of GWAS-associated lincRNAs may be nonconserved depending on the definition applied. Given these findings, computational, high-throughput functional and human pathophysiological approaches,[1, 4, 5, 9] rather than traditional metrics of conservation, should be applied to prioritize lncRNAs for functional studies. Expansion of research strategies using nontraditional model systems is urgently required to address physiological and pathophysiological functions of nonconserved lncRNAs and other genomic elements in human cardiometabolic disorders.

ARTICLE INFORMATION

Received July 15, 2020; accepted October 21, 2020.

Affiliations

Biostatistics Center, Massachusetts General Hospital, Boston (A.S.F., C.S., T.C.). Department of Medicine, Harvard Medical School, Boston, MA (A.S.F.). Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA (A.S.F., T.C.). Data Science Institute, Stanford University, CA (P.M.). Cardiology Division, Department of Medicine (M.E.O., E.C., H.L., C.X., M.P.R.) and Irving Institute for Clinical and Translational Sciences (M.P.R.), Columbia University, New York, NY.

Acknowledgments

We remember Xuan Zhang, a friend and outstanding scientist, who tragically and suddenly passed away while contributing to an early draft of this article.

Sources of Funding

Support for this research was provided by grants from the National Institutes of Health R01 GM127862 and R15 GM126485 to A.S. Foulkes and R01 HL132561, R01 HL113147, and K24 HL107643 to M.P. Reilly.

Disclosures

None.

Supplemental Materials

Tables I–V
Figures I–VI

REFERENCES

1. de Goede OM, Ferraro NM, Nachun DC, Rao AS, Aguet F, Barbeira AN, et al. Long non-coding RNA gene regulation and trait associations across human tissues. *bioRxiv*. Preprint posted online October 4, 2019. doi: 10.1101/793091
2. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep*. 2015;11:1110–1122. doi: 10.1016/j.celrep.2015.04.023
3. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*. 2014;505:635–640. doi: 10.1038/nature12943
4. Gil N, Ulitsky I. Regulation of gene expression by cis-acting long non-coding RNAs. *Nat Rev Genet*. 2020;21:102–117. doi: 10.1038/s41576-019-0184-5
5. Zhang X, Li DY, Reilly MP. Long intergenic noncoding RNAs in cardiovascular diseases: challenges and strategies for physiological studies and translation. *Atherosclerosis*. 2019;281:180–188. doi: 10.1016/j.atherosclerosis.2018.09.040
6. Shan K, Jiang Q, Wang XQ, Wang YN, Yang H, Yao MD, Liu C, Li XM, Yao J, Liu B, et al. Role of long non-coding RNA-RNCR3 in atherosclerosis-related vascular dysfunction. *Cell Death Dis*. 2016;7:e2248. doi: 10.1038/cddis.2016.145
7. Sallam T, Jones M, Thomas BJ, Wu X, Gilliland T, Qian K, Eskin A, Casero D, Zhang Z, Sandhu J, et al. Transcriptional regulation of macrophage cholesterol efflux and atherogenesis by a long noncoding RNA. *Nat Med*. 2018;24:304–312. doi: 10.1038/nm.4479
8. Michalik KM, You X, Manavski Y, Doddaballapur A, Zörnig M, Braun T, John D, Ponomareva Y, Chen W, Uchida S, et al. Long noncoding RNA MALAT1 regulates endothelial cell function and vessel growth. *Circ Res*. 2014;114:1389–1397. doi: 10.1161/CIRCRESAHA.114.303265
9. Mattioli K, Volders PJ, Gerhardinger C, Lee JC, Maass PG, Melé M, Rinn JL. High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. *Genome Res*. 2019;29:344–355. doi: 10.1101/gr.242222.118
10. Li C, Hu Z, Zhang W, Yu J, Yang Y, Xu Z, Luo H, Liu X, Liu Y, Chen C, et al. Regulation of cholesterol homeostasis by a novel long non-coding RNA LASER. *Sci Rep*. 2019;9:7693. doi: 10.1038/s41598-019-44195-2
11. Inouye M, Ripatti S, Kettunen J, Lyytikäinen LP, Oksala N, Laurila PP, Kangas AJ, Soininen P, Savolainen MJ, Viikari J, et al. Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS Genet*. 2012;8:e1002907. doi: 10.1371/journal.pgen.1002907
12. Holdt LM, Teupser D. Long noncoding RNA ANRIL: Lnc-ing genetic variation at the chromosome 9p21 locus to molecular mechanisms of atherosclerosis. *Front Cardiovasc Med*. 2018;5:145. doi: 10.3389/fcvm.2018.00145
13. Gao W, Zhu M, Wang H, Zhao S, Zhao D, Yang Y, Wang ZM, Wang F, Yang ZJ, Lu X, et al. Association of polymorphisms in long non-coding RNA H19 with coronary artery disease risk in a Chinese population. *Mutat Res*. 2015;772:15–22. doi: 10.1016/j.mrfmmm.2014.12.009
14. Congrains A, Kamide K, Oguro R, Yasuda O, Miyata K, Yamamoto E, Kawai T, Kusunoki H, Yamamoto H, Takeya Y, et al. Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B. *Atherosclerosis*. 2012;220:449–455. doi: 10.1016/j.atherosclerosis.2011.11.017
15. Ballantyne RL, Zhang X, Nuñez S, Xue C, Zhao W, Reed E, Salaheen D, Foulkes AS, Li M, Reilly MP. Genome-wide interrogation reveals hundreds of long intergenic noncoding RNAs that associate with cardiometabolic traits. *Hum Mol Genet*. 2016;25:3125–3141. doi: 10.1093/hmg/ddw154
16. Winkler TW, Justice AE, Graff M, Barata L, Feitosa MF, Chu S, Czajkowski J, Esko T, Fall T, Kilpeläinen TO, et al; arcOGEN Consortium; CHARGE Consortium; DIAGRAM Consortium; GLGC Consortium; Global-BPGen Consortium; ICBP Consortium; MAGIC Consortium. Correction: the influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study. *PLoS Genet*. 2016;12:e1006166. doi: 10.1371/journal.pgen.1006166
17. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Mägi R, Strawbridge RJ, Pers TH, Fischer K, Justice AE, et al; ADIPOGen Consortium; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GEFOS Consortium; GENIE Consortium; GLGC; ICBP; International Endogene Consortium; LifeLines Cohort Study; MAGIC Investigators; MuTHER Consortium; PAGE Consortium; ReproGen Consortium. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*. 2015;518:187–196. doi: 10.1038/nature14132
18. Pulit SL, Stoneman C, Morris AP, Wood AR, Glastonbury CA, Tyrrell J, Yengo L, Ferreira T, Marouli E, Ji Y, et al; GIANT Consortium. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum Mol Genet*. 2019;28:166–174. doi: 10.1093/hmg/ddy327

19. Rask-Andersen M, Karlsson T, Ek WE, Johansson Å. Genome-wide association study of body fat distribution identifies adiposity loci and sex-specific genetic effects. *Nat Commun*. 2019;10:339. doi: 10.1038/s41467-018-08000-4
20. Lumish HS, O'Reilly M, Reilly MP. Sex differences in genomic drivers of adipose distribution and related cardiometabolic disorders: opportunities for precision medicine. *Arterioscler Thromb Vasc Biol*. 2020;40:45–60. doi: 10.1161/ATVBAHA.119.313154
21. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z, et al; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; LifeLines Cohort Study. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*. 2014;46:1173–1186. doi: 10.1038/ng.3097
22. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, et al; Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013;45:1274–1283. doi: 10.1038/ng.2797
23. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D, Kyriakou T, Nelson CP, Hopewell JC, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet*. 2015;47:1121–1130. doi: 10.1038/ng.3396
24. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, Payne AJ, Steinthorsdottir V, Scott RA, Grarup N, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet*. 2018;50:1505–1513. doi: 10.1038/s41588-018-0241-6
25. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. 2019;47:D766–D773. doi: 10.1093/nar/gky955
26. Zhang X, Xue C, Lin J, Ferguson JF, Weiner A, Liu W, Han Y, Hinkle C, Li W, Jiang H, et al. Interrogation of nonconserved human adipose lincrnas identifies a regulatory role of linc-adal in adipocyte metabolism. *Sci Transl Med*. 2018;10:eaar5987. doi: 10.1126/scitranslmed.aar5987
27. Zhang H, Xue C, Wang Y, Shi J, Zhang X, Li W, Nunez S, Foulkes AS, Lin J, Hinkle CC, et al. Deep RNA sequencing uncovers a repertoire of human macrophage long intergenic noncoding rnas modulated by macrophage activation and associated with cardiometabolic diseases. *J Am Heart Assoc*. 2017;6:e007431. doi: 10.1161/JAHA.117.007431
28. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005;15:1034–1050. doi: 10.1101/gr.3715005
29. Nelson CP, Hamby SE, Saleheen D, Hopewell JC, Zeng L, Assimes TL, Kanoni S, Willenborg C, Burgess S, Amouyel P, et al; CARDIoGRAM+C4D Consortium. Genetically determined height and coronary artery disease. *N Engl J Med*. 2015;372:1608–1618. doi: 10.1056/NEJMoa1404881
30. Smit AFA, Hubley R, Green P. Repeatmasker open-3.0. http://www.repeatmasker.org. 1996-2010. February 12, 2020.
31. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4:44–57. doi: 10.1038/nprot.2008.211
32. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37:1–13. doi: 10.1093/nar/gkn923
33. Stender S, Smagris E, Lauridsen BK, Kofoed KF, Nordestgaard BG, Tybjaerg-Hansen A, Pennacchio LA, Dickel DE, Cohen JC, Hobbs HH. Relationship between genetic variation at PPP1R3B and levels of liver glycogen and triglyceride. *Hepatology*. 2018;67:2182–2195. doi: 10.1002/hep.29751
34. Zhang Z, Gao W, Long QQ, Zhang J, Li YF, Liu DC, Yan JJ, Yang ZJ, Wang LS. Increased plasma levels of lncRNA H19 and LIPCAR are associated with increased risk of coronary artery disease in a Chinese population. *Sci Rep*. 2017;7:7491. doi: 10.1038/s41598-017-07611-z
35. Karlsson T, Rask-Andersen M, Pan G, Höglund J, Wadelius C, Ek WE, Johansson Å. Contribution of genetics to visceral adiposity and its relation to cardiovascular and metabolic disease. *Nat Med*. 2019;25:1390–1395. doi: 10.1038/s41591-019-0563-7
36. Emdin CA, Khera AV, Natarajan P, Klarin D, Zekavat SM, Hsiao AJ, Kathiresan S. Genetic association of waist-to-hip ratio with cardiometabolic traits, type 2 diabetes, and coronary heart disease. *JAMA*. 2017;317:626–634. doi: 10.1001/jama.2016.21042
37. Ule J, Blencowe BJ. Alternative splicing regulatory networks: functions, mechanisms, and evolution. *Mol Cell*. 2019;76:329–345. doi: 10.1016/j.molcel.2019.09.017
38. Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet*. 2016;17:601–614. doi: 10.1038/nrg.2016.85
39. Ferguson JF, Xue C, Gao Y, Tian T, Shi J, Zhang X, Wang Y, Li YD, Wei Z, Li M, et al. Tissue-specific differential expression of novel genes and long intergenic noncoding RNAs in humans with extreme response to evoked endotoxemia. *Circ Genom Precis Med*. 2018;11:e001907. doi: 10.1161/CIRCGEN.117.001907
40. Zhang H, Shi J, Hachet MA, Xue C, Bauer RC, Jiang H, Li W, Tohyama J, Millar J, Billheimer J, et al. CRISPR/Cas9-mediated gene editing in human iPSC-derived macrophage reveals lysosomal acid lipase function in human macrophages-brief report. *Arterioscler Thromb Vasc Biol*. 2017;37:2156–2160. doi: 10.1161/ATVBAHA.117.310023
41. Rani N, Nowakowski TJ, Zhou H, Godshalk SE, Lisi V, Kriegstein AR, Kosik KS. A primate lncRNA mediates notch signaling during neuronal development by sequestering miRNA. *Neuron*. 2016;90:1174–1188. doi: 10.1016/j.neuron.2016.05.005
42. Durruthy-Durruthy J, Sebastiano V, Wossidlo M, Cepeda D, Cui J, Grow EJ, Davila J, Mall M, Wong WH, Wysocka J, et al. The primate-specific noncoding RNA HPAT5 regulates pluripotency during human preimplantation development and nuclear reprogramming. *Nat Genet*. 2016;48:44–52. doi: 10.1038/ng.3449
43. Xiao T, Liu L, Li H, Sun Y, Luo H, Li T, Wang S, Dalton S, Zhao RC, Chen R. Long noncoding RNA ADINR regulates adipogenesis by transcriptionally activating C/EBPα. *Stem Cell Reports*. 2015;5:856–865. doi: 10.1016/j.stemcr.2015.09.007
44. Rigoutsos I, Lee SK, Nam SY, Anfossi S, Pasculli B, Pichler M, Jing Y, Rodriguez-Aguayo C, Telonis AG, Rossi S, et al. N-BLR, a primate-specific non-coding transcript leads to colorectal cancer invasion and migration. *Genome Biol*. 2017;18:98. doi: 10.1186/s13059-017-1224-0
45. Yin X, Mead BE, Safaee H, Langer R, Karp JM, Levy O. Engineering Stem Cell Organoids. *Cell Stem Cell*. 2016;18:25–38. doi: 10.1016/j.stem.2015.12.005
46. Fatehullah A, Tan SH, Barker N. Organoids as an *in vitro* model of human development and disease. *Nat Cell Biol*. 2016;18:246–254. doi: 10.1038/ncb3312
47. Van Keuren ML, Gavrilina GB, Filipiak WE, Zeidler MG, Saunders TL. Generating transgenic mice from bacterial artificial chromosomes: transgenesis efficiency, integration and expression outcomes. *Transgenic Res*. 2009;18:769–785. doi: 10.1007/s11248-009-9271-2
48. Zhang Y, Pitchiaya S, Cieślik M, Niknafs YS, Tien JC, Hosono Y, Iyer MK, Yazdani S, Subramaniam S, Shukla SK, et al. Analysis of the androgen receptor-regulated lncRNA landscape identifies a role for ARLNC1 in prostate cancer progression. *Nat Genet*. 2018;50:814–824. doi: 10.1038/s41588-018-0120-1
49. Wang Y, Zeng X, Wang N, Zhao W, Zhang X, Teng S, Zhang Y, Lu Z. Long noncoding RNA DANCR, working as a competitive endogenous RNA, promotes ROCK1-mediated proliferation and metastasis via decoying of miR-335-5p and miR-1972 in osteosarcoma. *Mol Cancer*. 2018;17:89. doi: 10.1186/s12943-018-0837-6
50. Taylor J. Clues to function in gene deserts. *Trends Biotechnol*. 2005;23:269–271. doi: 10.1016/j.tibtech.2005.04.003
51. Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L. Evolution and functional classification of vertebrate gene deserts. *Genome Res*. 2005;15:137–145. doi: 10.1101/gr.3015505
52. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. Scanning human gene deserts for long-range enhancers. *Science*. 2003;302:413. doi: 10.1126/science.1088328
53. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature*. 2011;470:264–268. doi: 10.1038/nature09753
54. Huppi K, Pitt JJ, Wahlberg BM, Caplen NJ. The 8q24 gene desert: an oasis of non-coding transcriptional activity. *Front Genet*. 2012;3:69. doi: 10.3389/fgene.2012.00069
55. Volders PJ, Anckaert J, Verheggen K, Nuytens J, Martens L, Mestdagh P, Vandesompele J. LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res*. 2019;47:D135–D139. doi: 10.1093/nar/gky1031