Brief Communication:

# Functional disease architectures reveal unique biological role of transposable elements

Farhad Hormozdiari[*,1,2], Bryce van de Geijn[1,2], Joseph Nasser[2], Omer Weissbrod[1,2], Steven Gazal[1,2], Chelsea J.-T. Ju[3], Luke O'Connor[1,4], Margaux Louise Anna Hujoel[5], Jesse Engreitz[2], Fereydoun Hormozdiari[6,7], and Alkes L. Price [*1,2,5]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

[2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[3]Department of Computer Science, University of California, Los Angeles, California 90095

[4]Program in Bioinformatics and Integrative Genomics, Harvard Graduate School of Arts and Sciences, Boston, Massachusetts, USA

[5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

[6]Department of Biochemistry and Molecular Medicine, University of California, Davis, CA 95616, USA

[7]MIND Institute and UC-Davis Genome Center, Davis, CA 95616, USA

[*]Corresponding author: hormozdiari@hsph.harvard.edu, aprice@hsph.harvard.edu

**Abstract**

Transposable elements (TE) comprise roughly half of the human genome. Though initially derided as "junk DNA", they have been widely hypothesized to contribute to the evolution of gene regulation. However, the contribution of TE to the genetic architecture of diseases and complex traits remains unknown. Here, we analyze data from 41 independent diseases and complex traits (average $N$=320K) to draw three main conclusions. First, TE are uniquely informative for disease heritability. Despite overall depletion for heritability (54% of SNPs, 39±2% of heritability; enrichment of 0.72±0.03; 0.38-1.23 enrichment across four main TE classes), TE explain substantially more heritability than expected based on their depletion for known functional annotations (expected enrichment of 0.35±0.03; 2.11x ratio of true vs. expected enrichment). This implies that TE acquire function in ways that differ from known functional annotations. Second, older TE contribute more to disease heritability, consistent with acquiring biological function; SNPs inside the oldest 20% of TE explain 2.45x more heritability than SNPs inside the youngest 20% of TE. Third, Short Interspersed Nuclear Elements (SINE; one of the four main TE classes) are far more enriched for blood traits (2.05±0.30) than for other traits (0.96±0.09); this difference is far greater than expected based on the weaker depletion of SINEs for regulatory annotations in blood compared to other tissues. Our results elucidate the biological roles that TE play in the genetic architecture of diseases and complex traits.

# Introduction

Transposable elements (TE), defined as DNA sequences that can insert themselves at new genomic locations, comprise roughly half of the human genome[1,2]. Though initially derided as "junk DNA", TE have been widely hypothesized to contribute to the evolution of gene regulation by providing new targets for transcription factor binding and rewiring core regulatory networks[3–16]. TE have been shown to play these important roles in a growing number of specific examples, potentially impacting the genetic architecture of common disease. However, our current understanding of the contribution of TE to the genetic architecture of diseases and complex traits is extremely limited.

Here, we applied stratified LD score regression[17] (S-LDSC) with the baseline-LD model[18] to 41 independent diseases and complex traits (average $N$=320K) to estimate the components of heritability explained by different classes of TE. We sought to answer three questions. First, what is the contribution of TE to disease, and does this differ from what is expected based on the extent of their level of overlap with known functional annotations[19,20]? Second, do older TE contribute more to the disease heritability than younger TE? Third, do there exist classes of TE that play a greater role in specific diseases or traits?

# Results

## Overview of methods

We applied stratified LD score regression (S-LDSC)[17] to assess the contribution of different TE to disease and trait heritability. We estimated the heritability enrichment and standardized effect size ($\tau*$) for each TE annotation conditional on 75 functional annotations from the baseline-LD model[18] (Supplementary Table 1, see URLs). Heritability enrichment is defined as the proportion of heritability causally explained by the set of common SNPs in an annotation divided by the proportion of common SNPs in the annotation. Distinct from heritability enrichment, we also compute the enrichment that is expected based on an annotation's overlap with baseline-LD model annotations, denoted as Expected (baseline-LD) (see Methods); this computation determines the extent to which heritability enrichment/depletion is explained by known functional annotations. We note that enrichment and expected enrichment can be either $> 1$ or $< 1$ (i.e. depletion). Stan-

dardized effect size ($\tau*$) is defined as the proportionate change in per-SNP heritability associated with an increase in the value of the annotation by one standard deviation[18]; unlike heritability enrichment, $\tau*$ quantifies effects that are unique to the focal annotation (see Methods). For each TE annotation, we include an additional annotation defined by 500bp flanking regions, to guard against bias due to model misspecification[17] (see Methods). We have made our annotations and partitioned LD scores freely available (see URLs). Most of our results are meta-analyzed across 41 independent diseases and complex traits (Supplementary Table 2, same traits as in ref.[21]).

## TE are uniquely informative for disease heritability

We first focused on four main TE classes: long interspersed nuclear elements (LINE; 21% of SNPs), short interspersed nuclear elements (SINE; 16% of SNPs), long terminal repeats (LTR; 9.8% of SNPs), DNA transposons (DNA; 3.2% of SNPs), and the union all TE (ALLTE; 54% of SNPs). The proportion of SNPs in each TE class slightly exceeded the proportion of the genome spanned by the TE class (Supplementary Figure 1). This is consistent with weaker selective constraint within surviving TE, and confirms that SNPs lying inside TE can be effectively assayed despite the challenges of aligning TE sequences. ALLTE explained 39% of disease heritability (meta-analyzed across 41 diseases and traits), a moderate depletion (enrichment of 0.72±0.03; Figure 1A-B and Supplementary Table 3). The four main TE classes were all depleted or non-significantly enriched for trait heritability, with substantial heterogeneity between classes: 0.73±0.05 for LINE, 1.18±0.11 for SINE, 0.38±0.07 for LTR, and 1.23±0.19 for DNA (Figure 1B and Supplementary Table 3). Our simulations confirm that S-LDSC produces unbiased estimates of enrichment for these annotations (see Methods, Supplementary Figure 2). A secondary analysis of enrichment of fine-mapped causal disease SNPs[22,23] produced concordant results (Supplementary Table 4). A secondary analysis of enrichment of fine-mapped causal cis-eQTL SNPs[21] from GTEx data[24] also produced concordant results (Supplementary Table 5).

Notably, the heritability enrichments expected based on overlap with baseline-LD model annotations were much lower (Expected (baseline-LD); Figure 1A-B and Supplementary Table 3), consistent with the large depletion of overlap between TE and known functional annotations (Supplementary Figure 3). Accordingly, $\tau*$ estimates were significantly positive for each TE class (Figure 1C), implying disease heritability enrichment effects that are not captured by known func-

tional annotations. These $\tau*$ estimates were similar (in absolute value) to $\tau*$ estimates for the most informative annotations in our previous work[18,21]. Furthermore, each of the four main TE classes (LINE, SINE, LTR, and DNA) had a significant $\tau*$ conditional on each other and baseline-LD model annotations (Supplementary Table 6), indicating that each is uniquely informative for disease heritability.

We investigated whether the age of a TE impacts its contribution to disease heritability. We estimated the age of each TE using miliDiv (RepeatMasker software; see URLs), which computes the number of mutations relative to a consensus sequence to estimate the age of each TE[11]. We stratified SNPs lying in a TE into five quintiles based on the age of the TE. We determined that older SNPs had larger heritability enrichments than younger SNPs (e.g. $0.91\pm0.11$ for oldest quintile vs $0.37\pm0.10$ for youngest quintile; Supplementary Figure 4 and Supplementary Table 7). We repeated this analysis for each TE class (SINE, LINE, LTR, DNA) and observed the largest effect for SINE (Supplementary Figure 5). Analyses of Expected(baseline-LD) across TE families/subfamilies produced similar results, with the largest age effect for SINE (Supplementary Figure 6 and Supplementary Table 8). These results indicate that older TE have a higher contribution to disease heritability, perhaps because they have gained biological function.

Next, we analyzed 35 TE families/subfamilies spanning at least 0.4% of common SNPs (i.e., MAF $\geq$ 0.05; Supplementary Table 9). We identified 4 TE families/subfamilies that were significantly depleted for trait heritability (L1, L1PA3, ERV1, and L1PA4; Supplementary Figure 7 and Supplementary Tables 10 and 11); none were significantly enriched. For the 814 TE families/subfamilies spanning less than 0.4% of common SNPs (Supplementary Table 12), we estimated Expected (baseline-LD) enrichment only (Supplementary Table 13), as S-LDSC is not applicable to very small annotations[17]. We identified 587 TE families/subfamilies that were significantly depleted for expected disease heritability (Supplementary Table 14). We also identified 46 TE families/subfamilies that were significantly enriched for expected disease heritability (Supplementary Figure 8 and Supplementary Table 14), consistent with their excess overlap with known functional annotations (Supplementary Figures 9 and 10 and Supplementary Tables 15 and 16). Notably, LFSINE-Vert and AmnSINE1, which have previously been reported to have important biological function[25–27], had very large expected enrichments ($5.54\pm0.39$ and $5.44\pm0.32$ respectively).

## SINE are specifically strongly enriched for blood traits

We investigated whether TE enrichment varies across disease and traits. We estimated the heritability enrichment of each TE class (ALLTE, LINE, SINE, LTR, DNA) for 5 blood traits, 6 autoimmune diseases, and 8 brain-related traits (see Supplementary Table 17; same traits as in ref.[21]). We included a blood-specific chromatin annotation in our analyses of blood traits and autoimmune diseases, and a brain-specific chromatin annotation in our analyses of brain-related traits, in addition to the baseline-LD model (see Methods). Results are reported in Figure 2A and Supplementary Table 18. We determined that SINE are specifically strongly enriched for blood traits (2.05±0.30 vs. 1.18± 0.11 for non-blood traits; P=3E-04 for difference); no other TE class had significant trait class-specific enrichment after correcting for hypotheses tested, although SINE enrichment was non-significantly higher for autoimmune diseases vs. other traits (Supplementary Table 18). The difference in SINE enrichment for blood traits vs non-blood traits was much higher than expected based on overlap with baseline-LD model and blood-specific chromatin annotations (Expected (baselineLD+blood chromatin); Supplementary Table 18). Accordingly, we estimated a particularly large $\tau*$ for SINE for blood traits (Figure 2B and Supplementary Table 18), much larger (in absolute value) than $\tau*$ estimates for the most informative annotations in our previous work[18,21]. The specific importance of SINE for blood traits is consistent with the weaker depletion of SINE in blood-specific chromatin annotations vs. other tissue/cell types (Figure 2C and Supplementary Table 19), but is far greater than expected based on this weaker depletion; in particular, the $\tau*$ estimates of Figure 2B are conditioned on blood-specific chromatin annotations.

We repeated the trait class-specific analysis for the 35 TE families/subfamilies spanning at least 0.4% of common SNPs (Supplementary Tables 20-22). We did not detect any trait class-specific enrichments except for the Alu family, which spans 80% of the SINE class and produces results similar to SINE. For the 814 TE families/subfamilies spanning less than 0.4% of common SNPs, we detected 27 that had significantly higher Expected (baseline-LD+blood chromatin) enrichment for blood-related traits vs. other traits (Supplementary Table 23) and 27 that had significantly higher Expected (baseline-LD+blood chromatin) enrichment for autoimmune diseases vs other traits (Supplementary Table 24). The majority of TE families/subfamilies for that were specifically enriched for autoimmune diseases are endogenous retroviruses (ERV), including the MER41, which

has previously been reported to contribute to autoimmune disease[14]. We also detected 109 TE families/subfamilies with higher Expected (baselineLD+brain chromatin) enrichment for brain-related traits vs. other traits (Supplementary Table 25).

## Discussion

We have quantified the disease heritability explained by TE, including different classes of TE. We reached three main conclusions. First, TE are uniquely informative for disease heritability, as they explain substantially more heritability than expected based on their depletion for known functional annotations. This implies that TE acquire function in ways that differ from known functional annotations. Second, we observed that older TE contribute more to disease heritability, consistent with acquiring biological function. Third, the SINE class of TE is far more enriched for blood traits than for other traits, showing that TE biology can be trait class-specific.

Our findings have several biological implications. First, our results suggest that the functional annotation of the human genome is far from complete, as the functional regions underlying the contribution of TE to disease heritability have yet to be annotated. This motivates intense efforts to identify these functional regions. We have provided a framework ($\tau*$ metric; Figure 1C) to evaluate these efforts. Specifically, a $\tau*$ value close to 0 (conditional on a new set of functional annotations) would imply that this goal has been achieved; this can be evaluated for all TE and all traits (Figure 1C), but is of particular interest for SINE and blood traits (Figure 2B). Second, our TE-related annotations with conditionally significant signals (Figure 1C) can be incorporated to improve functionally informed fine-mapping[28–30], as well as functionally informed efforts to increase association power[31,32] and polygenic prediction accuracy[33–35].

We note several limitations of our work. First, S-LDSC cannot be applied to estimate the heritability enrichment of TE families/subfamilies that span a small proportion of the genome (e.g. less than 0.4% of common SNPs)[17]. We can instead compute the heritability enrichment that is expected based on an annotation's overlap with baseline-LD model annotations, although we caution that this quantity has a different interpretation. Second, we focused our analyses on common variants, as we used the 1000 Genomes LD reference panel, but future work could draw inferences about low-frequency variants using larger reference panels[36]. Third, SNPs lying inside TE may be

difficult to identify and annotate due to the challenges of aligning TE sequences. However, the proportion of SNPs in each TE class slightly exceeded the proportion of the genome spanned by the TE class (Supplementary Figure 1), suggesting that SNPs lying inside TE can be effectively assayed. In addition, we observed that 85% of 1000 Genomes SNPs lie in a 35-mer that has mappability of 1 (i.e. unique mappability) based on the ENCODE 35-mer track. Thus, 85% of the SNPs in our analysis are not impacted by aligning reads to repetitive regions of the genome. Furthermore, we confirmed that restricting our analyses to the 85% of SNPs with mappability of 1 produces very similar results (Supplementary Table 27). Fourth, inferences about components of heritability can potentially be biased by failure to account for LD-dependent architectures[18,37–39]. All of our analyses used the baseline-LD model, which includes 6 LD-related annotations[18]. The baseline-LD model is supported by formal model comparisons using likelihood and polygenic prediction methods, as well as analyses using a combined model incorporating alternative approaches[40]; however, there can be no guarantee that the baseline-LD model perfectly captures LD-dependent architectures. Despite these limitations, our results substantially improve our current understanding of the contribution of TE to the genetic architecture of diseases and complex traits.

## Acknowledgements

We are grateful to Armin Schoech and Po-Ru Loh for helpful discussions. This research was funded by NIH grants U01 HG009379, R01 MH101244, R01 MH109978 and R01 MH107649. This research was conducted using the UK Biobank Resource under Application 16549. F.H. is also supported by NIH grants T32 DK110919 and F32HG009987.

# Methods

## Heritability enrichment and standardized effect size ($\tau*$)

We use two metics (Heritability enrichment and standardized effect size ($\tau*$)) to measure the contribution of an annotation to disease and trait heritability[17,18]. We use S-LDSC to compute the heritability enrichment and standardized effect size ($\tau*$). S-LDSC assumes that the per-SNP heritability or variance of each SNP is equal to the linear contribution of each annotation[17]:

$$\mathrm{Var}(\beta_j) = \sum_c a_{cj}\tau_c \tag{1}$$

where $a_{cj}$ indicates the annotation value of SNP $j$ for the annotation $c$ and $\tau_c$ is the contribution of annotation $c$ to the per-SNP heritability. S-LDSC estimates the $\tau_c$ for each annotation using the following equation:

$$\mathrm{E}[\chi_j^2] = N\sum_c \ell(j,c)\tau_c + 1 \tag{2}$$

where $N$ is GWAS sample size and $\ell(j,c)$ is the LD-score for the SNP $j$ and annotation $c$ computed from the 1000 Genome project (see URLs). We estimated $\ell(j,c)$ as $\sum_k a_{ck}r_{jk}^2$ where $r_{jk}$ is the genotypic correlation between SNPs $j$ and $k$.

Because $\tau_c$ depends on trait heritability and the size of annotation we can not compare $\tau_c$ between different traits or annotations. Gazal et al.[18] introduced standardized effect size ($\tau*$) for an annotation as follows:

$$\tau_c* = \frac{\tau_c sd(c)}{h_g^2/M_c} \tag{3}$$

where $sd(c)$ is the standard deviation of the annotation values, $M_c$ is total number of common SNPs used to estimate the $h_g^2$, and $h_g^2$ is the SNP-heritability for each trait. In our experiments $M_c$ is equal to 5,961,159. We can compare $\tau*$ between different traits or annotations.

Heritability enrichment for an annotation is defined as the proportion of trait heritability captured by an annotation divide by the proportion of common SNPs that span that annotation. Thus,

heritability enrichment is computed as follows:

$$\text{Enrichment} = \frac{\%h_g^2(c)}{\%\text{SNP}(c)} = \frac{\frac{h_g^2(c)}{h_g^2}}{\frac{\sum_j a_{jc}}{M_c}} \tag{4}$$

where $h_g^2(c)$ is the heritability captured by the annotation $c$ and it is computed as follows:

$$h_g^2(c) = \sum_j a_{jc}\text{Var}(\beta_j) = \sum_j a_{jc}\Big(\sum_c a_{jc}\tau_c\Big) \tag{5}$$

Both heritability enrichment and $\tau*$ are computed conditional on set of annotations in the model (e.g. baseline-LD model[18]), $\tau*$ captures the signal that is unique to the focal annotation after conditioning on all the annotations in the model. Standardized effect size ($\tau*$) is defined as the proportionate change in per-SNP heritability associated with an increase in the value of the annotation by one standard deviation[18]. However, enrichment captures a signal that is unique and/or non-unique to the focal annotation.

We computed the statistical significance of heritability enrichment using block-jackknife, as described in our previous studies[17,18,21] where we break the genome to 200 equal blocks. We compute the statistical significant of $\tau*$ by assuming that $\frac{\tau*}{se(\tau*)}$ follows a normal distribution with mean zero and variance one ($\frac{\tau*}{se(\tau*)} \sim N(0,1)$)[17,18,21].

The meta-analyzed values of enrichment and $\tau*$ across the 41 independent traits (47 data sets, see Supplementary Table 2) were computed using a random-effect meta-analysis, as implemented in the `rmeta` R package (see URLs).

*Expected(baseline-LD):* We compute the expected (baseline-LD) enrichment for an annotation by assuming that the $\tau$ of the focal annotation is zero. This is equivalent to apply S-LDSC to each trait using baseline-LD model and compute the per-SNP heritability for each variant using equation (1). Next, we compute the $h_g^2(c)$ of the annotation by summing over the per-SNP heritability of all SNPs that are in the annotation $c$. In the end, we can compute the heritability enrichment using equation (4) and compute the standard error using similar block-jackknife.

## Proportion of heritability captured by each annotation

We can compute 3 different proportion of heritability (%heritability) for each annotation: observed heritability, Expected(%SNPs), Expected (baseline-LD). Observed heritability enrichment is obtained from S-LDSC by conditioning on baseline-LD model.

*Expected(%SNPs):* Under the null model, we assume the enrichment of an annotation is one. Thus, this annotation has none significant heritability enrichment (non-enriched or non-depleted), then we expect the %heritability for an annotation to be equal to %SNPs in that annotation.

*Expected(baseline-LD):* We compute this quantity by multiplying the expected (baseline-LD) and %SNPs.

*Observed %heritability:* We compute this quantity by utilizing S-LDSC results of heritability enrichment and %SNPs. We have: Observed %heritability = Heritability enrichment $\times$ %SNPs.

## TE annotations

We constructed two annotations for each TE where the first annotation is obtained by considering all the SNPs that fall in a TE and the second annotation is obtained by considering all the SNP in a 500bp window of the TE. The window annotation is based on recommendation of previous work[17]. All results are obtained by conditioning over baseline-LD model. The $\tau*$ and enrichment reported for each TE class/family/subfamily are based on the first constructed TE annotation. We compared this enrichment estimates with the case where we compute the enrichment of an annotation conditional jointly on 4 extra annotations created by considering different window size of 100, 200, 500, and 1000bp. We observed that S-LDSC results does not depend on the window size (Supplementary Figure 11).

## S-LDSC simulations

We set the $\tau$ for each annotation based on enrichment obtained in real data sets. Utilizing the total heritability we simulated causal trait effect sizes using a polygenetic model: $\beta \sim N(0, h_g^2/n_c)$ where $n_c$ is the number of causal SNPs. We simulated the phenotypic values under the additive model ($Y = X\beta + e$) where X is the standardized genotype matrix and $e$ is the environment and measurement noise. We computed the summary statistics by performing linear regression between

the phenotypic values and genotype data using PLINK software (see URLs). In our simulation, we vary the number of individuals for the traits among 2,000, 20,000, and 40,000 where UK biobank genotypes[41] are used. After simulating the summary statistics, we applied S-LDSC conditional on baseline-LD model and our TE annotation. Regression SNPs in S-LDSC were obtained from the HapMap Project phase 3[42] (see URLs). These SNPs are well-imputed SNPs. SNPs with marginal association statistics larger than 80 or larger than 0.001N and SNPs that are in the major histocompatibility complex (MHC) region were excluded from all the analyses[17,18,21]. Reference SNPs were obtained using the European samples in 1000G[41]. Heritability SNPs, which are used to estimate $h_g^2$, were common variants (MAF$\geq$0.05) in the set of reference SNPs.

## Excess overlap

Let $A$ and $B$ indicate two annotations and $|.|$ indicate the number of SNPs in the annotation. We defined the excess overlap as follows:

$$\text{Excess(A,B)} = \frac{\frac{|A \cap B|}{M}}{\frac{|A|}{M}\frac{|B|}{M}} \tag{6}$$

where $M$ is total number of SNPs and $|A \cap B|$ indicates the set of SNPs that is shared in both annotations $A$ and $B$. We compute the standard error over our estimates using block jackknife with 200 blocks that is similar how S-LDSC computed the standard error over heritability enrichment as described in our previous studies[17,18,21].

## Tissue-specific chromatin annotations

Blood chromatin is blood active chromatin regions by combining 27 blood cells and 6 chromatin marks (H3K27ac, H3K4me3, DNase, DNase-H3K27ac, DNase-H3K4me3) obtained from ChromImpute[43] applied on Roadmap Epigenomics data[20]. Non-blood chromatin is non-blood active chromatin regions by combining 100 non-blood cells and 6 chromatin marks (H3K27ac, H3K4me3, DNase, DNase-H3K27ac, DNase-H3K4me3).

Brain chromatin is brain active chromatin regions by combining 13 brain cells and 6 chromatin marks (H3K27ac, H3K4me3, DNase, DNase-H3K27ac, DNase-H3K4me3) obtained from ChromImpute[43] applied on Roadmap Epigenomics data[20]. Non-brain chromatin is non-brain active chro-

matin regions by combining 114 non-brain cells and 6 chromatin marks.

## URLs

baselineLD annotations: `https://data.broadinstitute.org/alkesgroup/LDSCORE/`

TE annotations: `https://data.broadinstitute.org/alkesgroup/LDSCORE/TE/`

Repeat masker software: `http://www.repeatmasker.org`

1000 Genomes Project Phase 3 data: `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502`

PLINK software: `https://www.cog-genomics.org/plink2`

BOLT-LMM software: `https://data.broadinstitute.org/alkesgroup/BOLT-LMM`

BOLT-LMM summary statistics for UK Biobank traits: `https://data.broadinstitute.org/alkesgroup/UK`

UK Biobank: `http://www.ukbiobank.ac.uk/`

UK Biobank Genotyping and QC Documentation: `http://www.ukbiobank.ac.uk/wp-content/uploads/2014`

`04/UKBiobank_genotyping_QC_documentation-web.pdf`

`rmeta` R package: `https://cran.r-project.org/web/packages/rmeta/index.html`

## References

1. McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences *36*, 344–355.

2. McClintock, B. (1984). The significance of responses of the genome to challenge. Science *226*, 792–801.

3. Kazazian, H. H. (2004). Mobile elements: Drivers of genome evolution. Science *303*, 1626–1632.

4. Biémont, C. and Vieira, C. (2006). Genetics: Junk DNA as an evolutionary force. Nature *443*, 521–524.

5. Slotkin, R. K. and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics *8*, 272–285.

6. Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics *9*, 397–405.

7. Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nature Genetics *42*, 631–634.

8. Lynch, V. J., Leclerc, R. D., May, G., and Wagner, G. P. (2011). Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nature Genetics *43*, 1154–1159.

9. Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, Â., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell *148*, 335–348.

10. Xie, M., Hong, C., Zhang, B., Lowdon, R. F., Xing, X., Li, D., Zhou, X., Lee, H. J., Maire, C. L., Ligon, K. L., et al. (2013). DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nature Genetics *45*, 836–841.

11. Jacques, P.-É., Jeyakani, J., and Bourque, G. (2013). The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genetics *9*, e1003504.

12. Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M. P., and Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Research *24*, 1963–1976.

13. Lynch, V. J., Nnamani, M. C., Kapusta, A., Brayer, K., Plaza, S. L., Mazur, E. C., Emera, D., Sheikh, S. Z., Grützner, F., Bauersachs, S., et al. (2015). Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. Cell Reports *10*, 551–561.

14. Chuong, E. B., Elde, N. C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087.

15. Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews Genetics *18*, 71–86.

16. Trizzino, M., Park, Y., Holsbach-Beltrame, M., Aracena, K., Mika, K., Caliskan, M., Perry, G. H., Lynch, V. J., and Brown, C. D. (2017). Transposable elements are the primary source of novelty in primate gene regulation. Genome Research *27*, 1623–1633.

17. Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics *47*, 1228–1235.

18. Gazal, S., Finucane, H. K., Furlotte, N. A., Loh, P.-R., Palamara, P. F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B. M., Gusev, A., et al. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature Genetics *49*, 1421–1427.

19. Consortium, T. E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

20. Kundaje, A., , Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

21. Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H. K., Ju, C. J.-T., Loh, P.-R., Schoech, A., Reshef, Y., Liu, X., O'Connor, L., et al. (2018). Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. Nature Genetics *50*, 1041–1047.

22. Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., et al. (2014). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.

23. Huang, H., Fang, M., Jostins, L., Mirkov, M. U., Boucher, G., Anderson, C. A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., et al. (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature *547*, 173–178.

24. Aguet, F., Brown, A. A., Castel, S. E., Davis, J. R., He, Y., Jo, B., Mohammadi, P., Park, Y., Parsana, P., Segrè, A. V., et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

25. Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., Rubin, E. M., Kent, W. J., and Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature *441*, 87–90.

26. Nishihara, H. (2006). Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Research *16*, 864–874.

27. Nishihara, H., Kobayashi, N., Kimura-Yoshida, C., Yan, K., Bormuth, O., Ding, Q., Nakanishi, A., Sasaki, T., Hirakawa, M., Sumiyama, K., et al. (2016). Coordinately co-opted multiple transposable elements constitute an enhancer for wnt5a expression in the mammalian secondary palate. PLOS Genetics *12*, e1006380.

28. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., and Raychaudhuri, S.

(2012). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nature Genetics *45*, 124–130.

29. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genetics *10*, e1004722.

30. Chen, W., McDonnell, S. K., Thibodeau, S. N., Tillmans, L. S., and Schaid, D. J. (2016). Incorporating functional annotations for fine-mapping causal variants in a bayesian framework using summary statistics. Genetics *204*, 933–958.

31. Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. The American Journal of Human Genetics *94*, 559–573.

32. Sveinbjornsson, G., Albrechtsen, A., Zink, F., Gudjonsson, S. A., Oddson, A., Másson, G., Holm, H., Kong, A., Thorsteinsdottir, U., Sulem, P., et al. (2016). Weighting sequence variants based on their annotation increases power of whole-genome association studies. Nature Genetics *48*, 314–317.

33. Shi, J., Park, J.-H., Duan, J., Berndt, S. T., Moy, W., Yu, K., Song, L., Wheeler, W., Hua, X., Silverman, D., et al. (2016). Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. PLOS Genetics *12*, e1006493.

34. Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X., and Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. PLOS Computational Biology *13*, e1005589.

35. Marquez-Luna, C., Gazal, S., Loh, P.-R., Furlotte, N., Auton, A., and and, A. L. P. (2018). Modeling functional enrichment improves polygenic prediction accuracy in UK biobank and 23andme data sets.

36. Gazal, S., Loh, P.-R., Finucane, H. K., Ganna, A., Schoech, A., Sunyaev, S., and Price, A. L. (2018). Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. Nature Genetics.

37. Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. The American Journal of Human Genetics *91*, 1011–1021.

38. Yang, J., , Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., Robinson, M. R., Perry, J. R. B., Nolte, I. M., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nature Genetics *47*, 1114–1120.

39. Speed, D., Cai, N., Johnson, M. R., Nejentsev, S., and Balding, D. J. (2017). Reevaluation of SNP heritability in complex human traits. Nature Genetics *49*, 986–992.

40. Gazal, S., Marquez-Luna, C., Finucane, H. K., and Price, A. L. (2018). Reconciling S-LDSC and LDAK models and functional enrichment estimates. biorxiv.

41. Durbin, R. M., Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., Vega, F. M. D. L., Donnelly, P., et al. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

42. Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

43. Ernst, J. and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nature Biotechnology *33*, 364–376.

44. Consortium, E. P. (2012). An integrated encyclopedia of dna elements in the human genome. Nature *489*, 57–74.

45. Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., et al. (2012). Integrative annotation of chromatin elements from encode data. Nucleic Acids Research *93*, 779–797.

46. Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., and Young, R. A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

47. Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482.

48. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

49. Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using gerp++. PLoS Computational Biology *6*, e1001025.

50. Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., et al. (2015). Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nature Genetics *47*, 1385–1392.

51. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., and Price, A. L. (2018). Mixed-model association for biobank-scale datasets. Nature Genetics *50*, 906–908.

# Figures

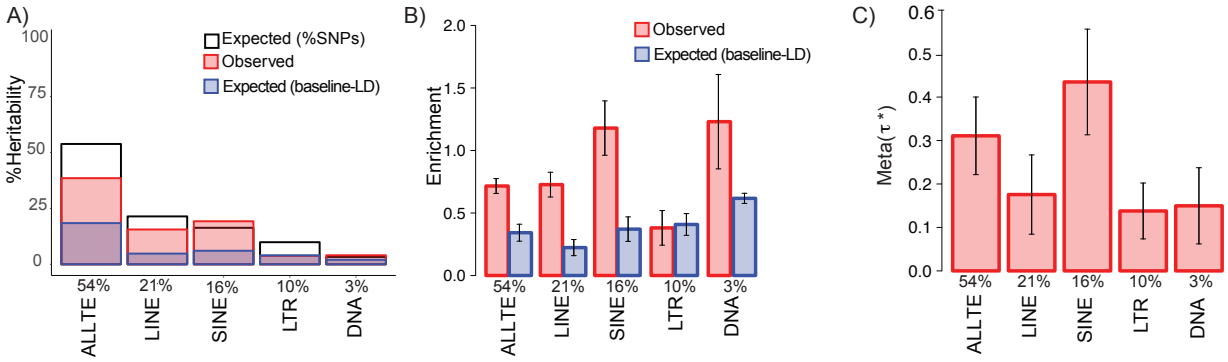



**Figure 1. TE are uniquely informative for disease heritability.** For each of four main TE classes and ALLTE, we report A) three measures of %heritability: Expected (%SNPs), Observed, and Expected (baseline-LD); B) two measures of heritability enrichment: Observed and Expected (baseline-LD); and C) standardized effect size ($\tau*$), which quantifies effect that are unique to the focal annotation. Results are meta-analyzed across 41 independent traits. Numerical values of %SNPs are provided for each annotation. Error bars denote 95% confidence intervals. Numerical results are reported in Supplementary Table 3.

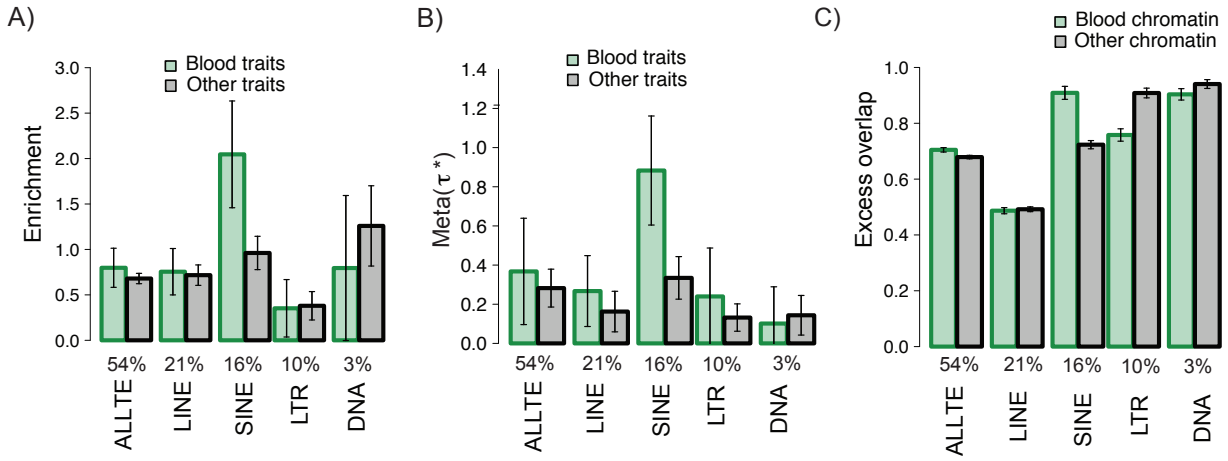
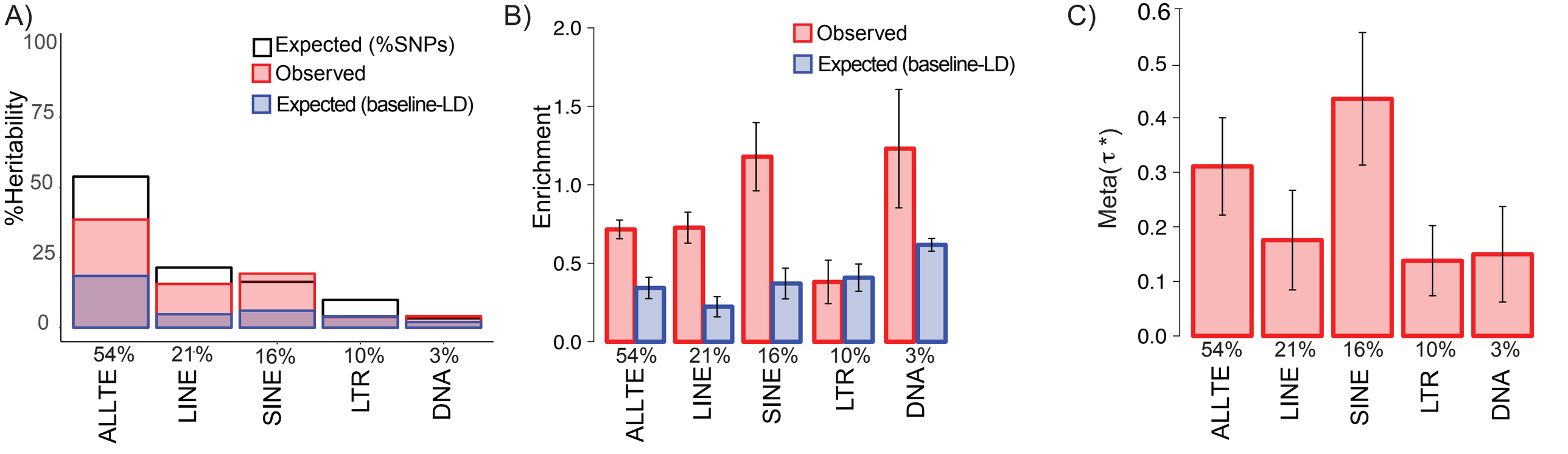
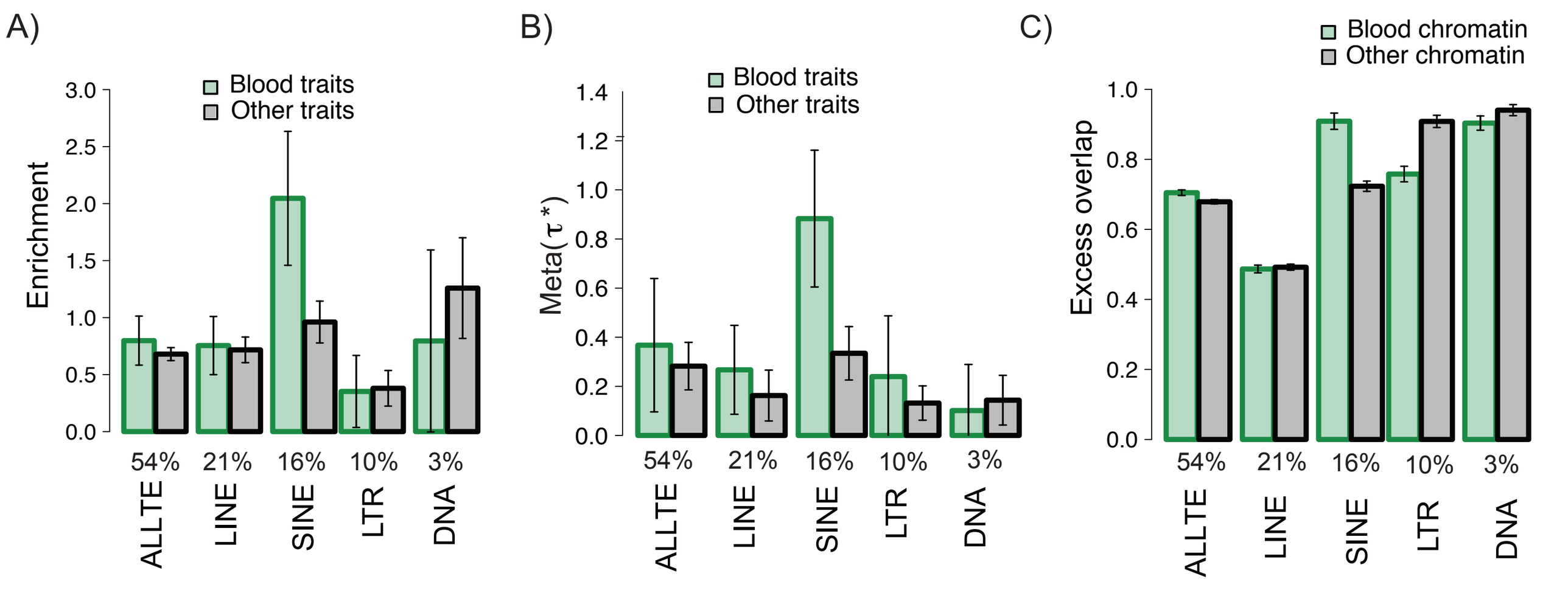



**Figure 2. Larger SINE enrichments for blood traits.** For each of four main TE classes and ALLTE, we report A) heritability enrichment for blood traits and other traits; B) standardized effect size ($\tau*$) for blood traits and other traits; and C) excess overlap with chromatin annotations in blood and chromatin annotations in other tissues. Results are meta-analyzed across 41 independent traits. Numerical values of %SNPs are provided for each annotation. Error bars denote 95% confidence intervals. Numerical results are reported in Supplementary Table 18 (A,B) and Supplementary Table 19 (C).


A)
%Heritability
100
75
50
25
0
Expected (%SNPs)
Observed
Expected (baseline-LD)
54%
21%
16%
10%
3%
ALLTE
LINE
SINE
LTR
DNA
B)
Enrichment
2.0
1.5
1.0
0.5
0.0
Observed
Expected (baseline-LD)
54%
21%
16%
10%
3%
ALLTE
LINE
SINE
LTR
DNA
C)
Meta(τ*)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
54%
21%
16%
10%
3%
ALLTE
LINE
SINE
LTR
DNA


A)
Blood traits
Other traits
Enrichment
3.0
2.5
2.0
1.5
1.0
0.5
0.0
54%
21%
16%
10%
3%
ALLTE
LINE
SINE
LTR
DNA
B)
Blood traits
Other traits
Meta(τ*)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
54%
21%
16%
10%
3%
ALLTE
LINE
SINE
LTR
DNA
C)
Blood chromatin
Other chromatin
Excess overlap
1.0
0.8
0.6
0.4
0.2
0.0
54%
21%
16%
10%
3%
ALLTE
LINE
SINE
LTR
DNA